

GENDER AND RELATIVE-AGE DIFFERENCES IN MATH FLUENCY  
USING CURRICULUM-BASED MEASUREMENT

by

Bonnie-Jean A. Foulds

B.Sc., University of British Columbia, 1978

Dip. in Learning Disorders, University of British Columbia, 1979

Ed. Cert., 1979

THESIS SUBMITTED IN PARTIAL FULFILMENT OF  
THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF EDUCATION

in

CURRICULUM AND INSTRUCTION

© Bonnie-Jean A. Foulds, 2002

THE UNIVERSITY OF NORTHERN BRITISH COLUMBIA

April, 2002

All rights reserved. This work may not be  
reproduced in whole or in part, by photocopy  
or other means, without the permission of the author.



National Library  
of Canada

Acquisitions and  
Bibliographic Services

395 Wellington Street  
Ottawa ON K1A 0N4  
Canada

Bibliothèque nationale  
du Canada

Acquisitions et  
services bibliographiques

395, rue Wellington  
Ottawa ON K1A 0N4  
Canada

Your file Votre référence

Our file Notre référence

**The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.**

**The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.**

**L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.**

**L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.**

0-612-80683-9

**Canada**


## APPROVAL

Name: Bonnie-Jean A. Foulds


Degree: Master of Education

Thesis Title: GENDER AND RELATIVE-AGE DIFFERENCES IN  
MATH FLUENCY USING CURRICULUM-BASED  
MEASUREMENT

Examining Committee:



Chair: Dr. Glenda Prkachin  
Associate Professor, Psychology Program  
UNBC



Supervisor: Dr. Peter MacMillan  
Assistant Professor, Education Program  
UNBC



Committee Member: Dr. Dennis L. C. Procter  
Assistant Professor, Education Program  
UNBC



Committee Member: Colin Chasteauneuf, BEd, MEd  
Instructor, Education Program  
UNBC



External Examiner: Fred Egglestone, EdD, BC Ed Cert  
Principal, Prince George Secondary School

Date Approved:

April 26, 2002

## ABSTRACT

This research is comprised of several studies using Curriculum-Based Measurement (CBM) math fluency from a Canadian school district. The pre-study investigated inter-rater reliability for 38 markers. After marking 14 Grade 7 CBM math probes a correlation mean of .98, and median of .99 were calculated for the markers. Unanimous agreement was reached by the markers on 40% of the questions completed by the students. Despite a high correlation between markers additional analysis determined several issues contributing to marker discrepancy including addition errors, unmarked questions, and not following marking rules.

The main study investigated gender and relative-age differences, effect sizes, and effect size comparisons. Two additional studies examined performance group differences and grade retention. A sample consisting of 1754 Grade 1 to 7 students participated in all aspects of the main study and performance group differences. Seventy Grades 1 to 7 students, eliminated from the main study, comprised the grade retention study. Using a  $2 \times 3 \times 3$  repeated measures ANOVA the results indicated no evidence of gender, relative-age differences or an interaction. Additional investigation with *t* tests, and effect sizes noted a gender difference for Grade 1 and 2 only. Most effect sizes were trivial for gender and relative-age differences. One gender or relative-age was not favoured over another. Effect size results were compared to CBM reading and writing as well as the results of other math research. Differences between performance groups were not evident. Retained students did not perform as well as the students who were in the appropriate grade for their age.

## TABLE OF CONTENTS

<i>Abstract .....</i>	<i>ii</i>
<i>Table of Contents .....</i>	<i>iii</i>
<i>List of Tables .....</i>	<i>vi</i>
<i>Acknowledgments .....</i>	<i>vii</i>
<b>CHAPTER ONE: INTRODUCTION .....</b>	<b>1</b>
<i>Curriculum-Based Measurement .....</i>	<i>4</i>
<i>Description of the School District .....</i>	<i>5</i>
<i>History of Curriculum-Based Measurement in the School District .....</i>	<i>6</i>
<b>Research Problem and Hypotheses .....</b>	<b>9</b>
<b>Definition of terms .....</b>	<b>11</b>
<i>Research Questions .....</i>	<i>13</i>
<i>Hypotheses .....</i>	<i>14</i>
<b>CHAPTER TWO: LITERATURE REVIEW .....</b>	<b>17</b>
<b>Curriculum-Based Measurement .....</b>	<b>17</b>
<i>Development and Use of Curriculum-Based Measurement .....</i>	<i>17</i>
<i>Limitations of Curriculum-Based Measurement .....</i>	<i>19</i>
<i>Inter-rater reliability of Curriculum-Based Measurement .....</i>	<i>21</i>
<i>Curriculum-Based Measurement Research in School District No. 57 .....</i>	<i>24</i>
<b>Research in Mathematics Achievement .....</b>	<b>27</b>
<i>Gender Differences in Math .....</i>	<i>27</i>
<i>Gender differences favouring girls .....</i>	<i>30</i>
<i>No evidence of gender differences .....</i>	<i>31</i>
<i>Gender differences in the high performance group .....</i>	<i>33</i>
<i>Relative-age Differences .....</i>	<i>37</i>
<i>Relative-age differences in the high performance group .....</i>	<i>40</i>
<i>Relative-age differences and science .....</i>	<i>41</i>
<i>Relative-age differences and sports .....</i>	<i>41</i>
<i>No evidence of relative-age differences .....</i>	<i>42</i>
<i>Gender and Relative-Age Differences Using Curriculum-Based Measurement Math .....</i>	<i>45</i>
<i>Grade Retention .....</i>	<i>46</i>
<b>Significance of the Proposed Study .....</b>	<b>47</b>
<b>CHAPTER THREE: DESIGN AND METHODS .....</b>	<b>50</b>
<b>Subjects .....</b>	<b>50</b>
<i>Inter-Rater Reliability .....</i>	<i>50</i>
<i>Gender and Relative-Age Study .....</i>	<i>51</i>
<i>Data cleaning and screening .....</i>	<i>54</i>
<i>Performance Group Study .....</i>	<i>55</i>
<i>Grade Retention Study .....</i>	<i>55</i>
<b>Instrumentation .....</b>	<b>56</b>
<i>Inter-Rater Reliability .....</i>	<i>56</i>
<i>Gender and Relative-Age Study .....</i>	<i>56</i>
<b>Procedures .....</b>	<b>57</b>
<i>Inter-Rater Study .....</i>	<i>57</i>
<i>Gender and Relative Age Study .....</i>	<i>58</i>
<i>Performance Group Study .....</i>	<i>60</i>
<i>Grade Retention Study .....</i>	<i>60</i>

<b>Data Analysis.....</b>	<b>61</b>
<i>Inter-Rater Reliability.....</i>	<i>61</i>
<i>Gender and Relative-Age Differences .....</i>	<i>62</i>
<i>Considerations for repeated measures analysis .....</i>	<i>62</i>
<i>Effect size comparisons.....</i>	<i>63</i>
<i>Performance Group Study .....</i>	<i>63</i>
<i>Grade Retention Study.....</i>	<i>64</i>
<b>Ethics .....</b>	<b>64</b>
<b>CHAPTER FOUR: RESULTS.....</b>	<b>66</b>
<b>Inter-Rater Reliability Findings.....</b>	<b>66</b>
<i>Reasons for Differences in Probe Scores.....</i>	<i>76</i>
<i>Addition Errors .....</i>	<i>76</i>
<i>Question Score Discrepancies .....</i>	<i>79</i>
<i>Unmarked questions .....</i>	<i>79</i>
<i>Differences in question scores between markers .....</i>	<i>80</i>
<b>The Main Studies .....</b>	<b>84</b>
<i>Gender and Relative-Age Differences .....</i>	<i>84</i>
<i>Means, Mean Square Within and Effect Sizes.....</i>	<i>88</i>
<i>Effect Size Comparisons .....</i>	<i>94</i>
<i>Comparison with CBM reading and writing.....</i>	<i>94</i>
<i>Comparison with other math research .....</i>	<i>97</i>
<i>Performance Group Differences.....</i>	<i>99</i>
<i>Grade Retention.....</i>	<i>104</i>
<b>CHAPTER FIVE: DISCUSSION AND CONCLUSIONS .....</b>	<b>108</b>
<b>Inter-Rater Reliability.....</b>	<b>108</b>
<i>Marking Considerations .....</i>	<i>109</i>
<i>Addition Errors .....</i>	<i>109</i>
<i>Question Score Discrepancy.....</i>	<i>110</i>
<i>Unmarked questions .....</i>	<i>111</i>
<i>Rules not followed.....</i>	<i>111</i>
<i>Additional rules .....</i>	<i>111</i>
<i>Other marking issues .....</i>	<i>112</i>
<i>Impact of Marker Differences.....</i>	<i>113</i>
<i>Recommendations to Reduce Marker Differences .....</i>	<i>114</i>
<b>The Main Study .....</b>	<b>116</b>
<i>Gender and Relative Age Differences.....</i>	<i>117</i>
<i>Gender differences.....</i>	<i>117</i>
<i>Relative-age differences.....</i>	<i>120</i>
<i>Gender and relative-age interactions .....</i>	<i>121</i>
<i>Effect size comparisons.....</i>	<i>122</i>
<i>Performance Group Differences.....</i>	<i>124</i>
<i>Grade Retention.....</i>	<i>126</i>
<b>Limitations of the Study.....</b>	<b>126</b>
<i>Inter-Rater Study .....</i>	<i>127</i>
<i>The Main Study.....</i>	<i>129</i>
<b>Implications for Future Study and Practice.....</b>	<b>130</b>
<b>REFERENCES.....</b>	<b>135</b>
<b>APPENDICES.....</b>	<b>143</b>
<b>Appendix A: Letters of Permission.....</b>	<b>144</b>

<i>Appendix B: Draft Technical Report of the CBM (Math) Norming Project.....</i>	<i>148</i>
<i>Appendix C: Draft Math Norms Tables for Curriculum Based Measurement Calculation.....</i>	<i>167</i>
<i>Appendix D: Sample CBM Math Probe and Answer Key.....</i>	<i>177</i>
<i>Appendix E: The CBM Math (Calculation) Norming Training Project, 1999-2000 Hand-outs.....</i>	<i>182</i>
<i>Appendix F: Inter-Rater Consent Form and Letter .....</i>	<i>199</i>
<i>Appendix G: Examples of Questions with Marker Discrepancies .....</i>	<i>202</i>
<i>Appendix H: Repeated-Measures ANOVA for the Average Performance Group.....</i>	<i>207</i>

## LIST OF TABLES

Table 1: Correct Digit Scores for Inter-Rater Reliability Probes.....	67
Table 2: Inter-Rater Correlation Coefficients .....	68
Table 3: Inter-Rater Means and Standard Deviations .....	72
Table 4: Probe Means and Standard Deviations Calculated by the Markers .....	75
Table 5: Impact of Addition Errors by Markers .....	77
Table 6: Discrepancies in Marking.....	80
Table 7: Causes of Marker Disagreement on Specific Questions.....	82
Table 8: Concerns Which Did Not Contribute to Marker Disagreement.....	83
Table 9: Repeated-Measures ANOVA for Gender and Relative-Age .....	85
Table 10: Mean CD Scores and Effect Sizes for Grade and Gender by Norming Period.....	89
Table 11: Mean CD Scores and Effect Sizes for Grade and Relative-Age by Norming Period .....	92
Table 12: CBM Math and Reading Gender Effect Size Comparison .....	95
Table 13: CBM Math and Writing Gender Effect Size Comparison .....	96
Table 14: Gender Effect Size Comparison for CBM Math and Other Math Research.....	98
Table 15: Repeated-Measures ANOVA for the Low Performance Groups.....	100
Table 16: Repeated-Measures ANOVA for the High Performance Groups.....	101
Table 17: Number of Students by Gender and Performance Group .....	104
Table 18: Mean CD for Retained and Appropriate Age Students by Grade and Gender .....	105



## ACKNOWLEDGEMENTS

I would like to express my appreciation and thanks to Dr. Peter MacMillan for his guidance, support, and assistance in the development and completion of this thesis.

A thank-you to all of my family for the support and encouragement they have shown since I began the masters program. My husband Terry, and my children Heather-Jean and Gordon demonstrated their patience while inspiring me to complete this endeavour.

I also owe thanks to many friends and colleagues who assisted, and encouraged me while I was completing this thesis.

## CHAPTER ONE: INTRODUCTION

Academic achievement is a continual topic of concern. Accountability is often considered an issue in education. Comparison of students with each other, and against the perceived or determined “average”, is deemed necessary to provide accountability. Through performance or achievement comparisons, it is possible to determine if students are making progress relative to themselves or their peers. Student comparison happens in classrooms, schools, and districts and even among students from different countries. Frequently, a concern regarding achievement centres on the way one subgroup fares in comparison to the total group average. The ongoing controversy regarding gender is whether boys and girls academic performance is similar, or whether one gender outperforms the other. Specifically, “concerns [exist] about the failure of female students to achieve their academic potential” (Wentzel, 1988, p. 691). Another debate raised is how the month of birth impacts a student’s academic performance. Gender issues are also surfacing regarding differences between boys and girls identified as the top performers in a class. Does gender, relative-age, or ability level impact student achievement? Further investigation is required to determine if gender, relative-age, or ability level are worth consideration.

The National Post Online on December 6<sup>th</sup>, 2000 reported a study in May 1999 that compared Canadian students to students from other countries in math and science (see Mullis, Martin, & Gonzalez, et al., 1999 for fuller details regarding the Third International Math and Science Study). For some time an ongoing concern in math has been the underachievement of females compared to males. This international study addressed the issue of Canadian students and the results of the study indicated that

“Canadian boys and girls did equally well in math, but boys outperformed girls by a substantial margin in science” (“Canadian students near top,” 2000). Another article discussed provincial test results of Ontario students and argued that, for gender differences, “the gap narrows significantly in math ... as children go through high school” (Lindgren, 2000). Both articles indicated that, despite the history of concern regarding gender differences in math, it is questionable whether there is still reason for concern.

Owens (2001) reported in a National Post Online article that boys are the disadvantaged gender when entering kindergarten. In contrast to “Canadian students near top” (2000) and Lindgren (2000), Owens suggested “boys should start kindergarten a year later than girls to compensate for their slower development rate.” This slower development of boys not only impacts their Kindergarten performance but, according to Owens, also impacts later development. He demonstrated this with Grade 3 and Grade 6 test results in Ontario which indicated that “a clear majority of boys do not reach the provincial standard in either reading or writing, while a clear majority of girls do, [whereas] in math boys trail girls by about 10%.” Owens recommends a method to solve relative-age concerns as well as retention issues. If younger students, born in the last three months before the school entry cut-off, develop more slowly than the average age, or their older peers of the same birth year and grade, then one can infer they will not perform academically as well their peers. Owens is claiming that retaining a student will increase a student’s chance to develop academically at the same level as their year younger peers. Thus, Owens suggests that younger or less academically developed

students will benefit from grade retention and should not be expected to perform as well as their age appropriate peers.

These articles suggest that the questions surrounding the existence of gender and relative-age differences do not have definitive answers. For educators, parents and students, the answers to concerns surrounding gender and relative-age differences in math have not been provided. In order to investigate these issues, a measurement tool is required. Many researchers, especially in the United States, use standardized tests to explore the existence of gender and relative-age differences (Hall, Davis, Bolen & Chia, 1999; Olson, 1989; Rabinowitz, 1989). However, a number of potential difficulties can arise when standardized tests are used. Test users must possess the qualifications to select, administer, score, and interpret the test (Sax & Newton, 1997).

One possible measurement tool researchers could use which is not a standardized test is Curriculum-Based Measurement (CBM). A Canadian school district is presently using CBM as an alternate measurement tool. Therefore, it is possible to investigate the existence of gender, relative-age, performance group, and grade retention differences in math without using a standardized test.

The assessment tool implemented to explore the existence of differences in gender, relative-age, performance group or grade retention, does not change the need for a response to the results. No evidence of differences may result in no action or further investigation. If differences are found, then educators must determine how they will attempt to correct these differences. One option could be to allow the differences to continue without intervention. Alternatively, the educators could pursue additional

research to determine possible causes of the differences. Finally, educators could implement strategies to change the existence of differences.

### *Curriculum-Based Measurement*

In response to the negative attitudes towards standardized testing, educators have developed alternative methods of evaluation (Daniels, 1999). These alternate assessments include Curriculum-Based Assessment (CBA) and CBM. Curriculum-Based Assessment allows educators to measure a student's growth on short-term objectives. As the student masters the curriculum, CBA changes the test format. The difficulty level of the test then increases (Fuchs & Fuchs, 1992; Deno, 1992). In contrast, CBM, which is a form of CBA, measures the student's progress toward a specified goal at the end of the school year. Therefore, the CBM test format and difficulty remain the same (Allinder & Eccarius, 1999; Fuchs & Fuchs, 1992).

Another difference between CBM and CBA "is that CBM employs a prescriptive set of measurement procedures with documented reliability and validity" (Fuchs & Fuchs, 1992, p. 45). According to Marston (1989), these measures or probes must be "(1) tied to the students' curricula, (2) of short duration to facilitate frequent administration, (3) capable of having multiple forms, (4) inexpensive, and (5) sensitive to the improvement of students' achievement over time" (p.30). The CBM tests used to measure academic progress are called probes. Probes measure basic academic skills in reading, writing, math, and spelling.

CBM was first introduced as a tool for making decisions regarding student progress and programming in special education (Deno, 1989; Fuchs, Fuchs, Hamlett, &

Stecker, 1990; Shinn, 1989). Since the development of CBM, it continues to be used as a tool to assist in making decisions regarding special education services in school districts.

### *Description of the School District*

One Canadian school district using CBM as an alternate assessment tool is School District No. 57 (SD57), Prince George. This district is located in the central interior of the province of British Columbia and is a large district both geographically and in student population. The school district is composed of the city of Prince George as the major centre, the three smaller communities of Mackenzie, McBride, and Valemount and the areas in between. As a district, it has schools located in a variety of settings including inner-city, suburban and rural. Inner-city schools have many students who live in poverty, come to school hungry, and who benefit from additional resources and programs. During the 1999-2000 school year 18,664 students were enrolled in the schools, of whom 10,872 were elementary students. Of the total number of students, 10 to 12% are self-identified as aboriginal. Status and Non-status, Metis, and Inuit are all considered aboriginal. The school district employed approximately 1170 teachers in 1999-2000.

Hedekar (1997) completed a similar study to this one in SD57. Although the academic subjects studied by Hedekar differed from mine, many other factors of her study resemble this one. Minor changes have taken place in the district since completion of Hedekar's study. The district has experienced a slight decrease in student and teaching population. Some staff turnover has taken place within the district. A few teachers have remained in the same teaching position since development of the first CBM norming project. This increased the consistency in administration and scoring of the CBM probes

between the first CBM norming project and the CBM math norming project. Thus, the data collected for this study comes from a similar student population as in the original CBM study in the school district.

#### *History of Curriculum-Based Measurement in the School District*

Educators in SD57, became interested in CBM in 1985 (School District No. 57, 1996). Teachers began implementing the CBM reading, spelling and written expression techniques as a means of monitoring student progress after attending a 1991 Tri-University Summer Institute in Curriculum Based Measurement held in Prince George. Subsequent research by Area Support Team members, combined with the confirmatory experiences of teachers who used CBM probes in their classrooms, led to the decision that CBM would support the principles, guidelines and recommendations of the School Support Services Task Force established in June, 1993. It was determined that the CBM procedures, which combined regular assessments with effective interventions, would fit with the new problem-solving model developed by the district. Consequently, the district began a research project into locally developed CBM norms, one that would assist educators using the School District problem-solving model.

The first CBM Research Project in SD57 began in Spring 1995, and was overseen by a joint committee of the School District and the University of Northern British Columbia (UNBC). A plan was formulated for the development of local CBM norms at the elementary level for reading and written expression (School District No. 57, 1996). Data for reading and written expression norms were collected during the 1995-1996 school year. Prior to data collection, a teacher representative from every elementary school in the school district attended a one-day training session to learn how to

administer and score the CBM probes. A total of 1849 students from Grades 1 to 7 participated in the project. Three norming periods for Grades 2 to 7 took place in October, January and April. Grade 1 students only participated in the April norming period. A stratified, random sample of approximately 20% of the student population in Grades 1 to 7 participated in this original SD57 CBM norming project. Following collection of the data by SD57, it was sent to the UNBC. At UNBC the data was analyzed, and Dr. Peter MacMillan developed norms for each grade level and norming period. The CBM norms for reading and written expression were then presented at an inservice held the following school year to school personnel in written form as a Guidebook.

Popular concern for student comparison had further implications for schools regarding the use of CBM. In SD57 the concern for student achievement lead to the formation of an Academic Achievement Committee in each school. The CBM reading, written expression and math norms developed by SD57 are elementary school indices available for setting academic achievement goals. In fact, almost all elementary schools in SD57 now use the reading CBM scores. Less than 50% of the elementary schools use the writing CBM scores to monitor the school's achievement against the district norms. About 50% of the elementary schools plan to use the CBM math scores to monitor progress as part of their Academic Achievement plan when the math norms became available. The opportunity to compare school norms to district norms enables schools to determine academic growth, allocate resources, and make program decisions which impacts students' future academic development.



Curriculum-Based Measurement was originally introduced into SD57 as a technique to monitor student progress. However, as outlined by Fewster (2000), it was also adopted by SD57 in 1994 to assist in identifying students requiring support as part of the “formal problem-solving model for the delivery of special education services” (p.2). At the time of this writing, CBM is considered a valuable tool in all elementary schools of the district, and is used to monitor student progress, identify students requiring special education assistance, and monitor student achievement. The work by Fewster showed that the reading and writing fluency norms predicted academic success in the humanities areas in secondary school. Consequently, while CBM probes and their related norming information are useful assessment tools for determining the current success of an elementary student, they also serve as predictors of future academic success in Grades 8, 9, and 10.

The use of CBM in schools is not limited to the reading and writing techniques taught for the first norming project. Additional inservices in SD57 introduced teachers to CBM techniques for monitoring spelling and additional writing assessment techniques. A few educators developed some of their own CBM measures to monitor student progress. Hence, many staff members, including this researcher, made requests for the development of further valid and reliable CBM math norms to provide additional tools for assessing and monitoring student progress. As a result, the district undertook a second major CBM norming project in math computation for students in Grades 1 to 7, during the 1999-2000 school year.

## Research Problem and Hypotheses

The existence of gender differences in the area of math continues to be a concern in schools. Historically, the focus of this concern has been on the under-achievement of girls in comparison to boys (Sadker, 1994). Several factors may contribute to gender differences. Nevertheless, it is difficult to know if there should be a concerted effort put into discovering the cause, without knowing the state of the gender differences.

Discovering if gender differences do exist will provide a first step to dealing with the cause. The existence of consistent gender differences throughout all the grades is unknown. It is difficult to determine if gender differences exist without consistent measures to provide a reliable and valid comparison.

A further concern in the area of math often discussed by parents and school staff is the relative-age (as determined by the month of birth within a specific calendar year) of students and their academic success. Hearing that a student has a “late” birthday (as defined by having a birthday in the last three months of a calendar year) often goes hand-in-hand with the expectation that the student may not be progressing or achieving at the same rate as other students. In contrast, it is expected that older students in the same grade, whose birthday is in the first three months of a calendar year, will be the high achievers in their grade. At present, evidence regarding the impact of relative-age on academic achievement is contradictory (Boyd, 1989; Olson, 1989).

If however, gender differences do not exist within the average population, researchers question the concern about gender differences between the high and low achievers. It is possible that more boys than girls are among the high achievers in the area of math. Gender differences within selected populations could impact future academic

and professional choices of students. Unless it can be determined that gender differences exist within high or low performance groups, their effect is unknown.

Finally, if a student has difficulty achieving academically, grade retention is considered a solution to increase the student's opportunity for academic success. It could be argued that academic success is more assured by holding students back one year so that they are among the older students in the class, rather than among the younger students in their present grade. Placing a student in a grade below what is appropriate for their age provides them with the opportunity to achieve academically as well as or better than their age appropriate peers do. What remains unknown is if retaining a student in a grade increases their chance for success in math.

To address the concerns raised by these questions, a method of reliably answering these questions must be available. Use of the data collected for the CBM math norming project in SD57, during the 1999-2000 school year is one possible way to answer these questions. However, since many different people undertook the scoring and administration of these probes, the question arises about the reliability of the marking. According to Hedekar (1997), the markers in her CBM study of reading and written expression produced reliable results. Before this present study, it was unknown if there was enough consistency between the markers of the CBM math data to consider the results reliable. It has not been determined if the CBM math norming data can reliably answer questions regarding the influence of gender, relative-age, performance group, and grade retention differences.

This researcher looked at marker reliability (inter-rater reliability) as it relates to the main study. The markers' reliability was determined before undertaking the other

studies regarding CBM math. The results of the Inter-Rater study add to the reliability of the CBM math gender and relative-age results. If marker reliability is low, the results for the rest of the study could be considered invalid.

### Definition of terms

Note that where definitions used in this study are taken directly from another reference source their quotation marks are eliminated. Other sources of definitions are acknowledged for each individual term in the parenthesis following the definition.

*Curriculum-Based Measurement (CBM)* refers to specific procedures used for measuring pupil proficiency within basic skills of the curricula. The basic skills typically measured are reading, written expression, math and spelling. CBM has documented reliability and validity. The skills assessed represent the curriculum for a complete school year and use year-end goals (Fuchs et al., 1990; Fuchs & Fuchs, 1992).

*CBM Math Data* refers to the analyzed CBM math scores acquired for all the students during one or more math probe administrations in a defined setting. The setting could be a classroom, school, or school district. For this study the setting is SD57.

*Norms* are scores determined for the students at each grade level, established through testing, against which subsequent testing can be analyzed. Elliot and Bretzing state norms are percentiles, or standard score conversions, derived from a distribution of scores earned by an identified group (cited in Hedekar, 1997, p. 23). In this study, CBM math norms refers to the norms Walraven and MacMillan (2000) developed for SD57 using the CBM math data collected during the 1999-2000 school year (see Appendix B and C).

*Probes* are concise CBM measurement tests designed to assess skills in reading, writing, spelling, and math fluency (Hedekar, 1997) and are relevant to the curriculum for the school year (Allinder & Eccarius, 1999; Deno, 1992; Howell, Fox & Morehead, 1993; Shinn et al. 1990). A math probe consists of two pages of representative math questions from the math curriculum for a specific grade. Each probe has 15 questions on each of the two pages. Math probes are administered for five minutes. For a sample CBM math probe and answer key see Appendix D.

*Correct digits (CD)* in math fluency, refers to credit earned for each digit that is correct within a student response (Baker, Collins, & Goodwin, 1992, p. 66).

*CD score* is the total number of correct digits earned on a CBM math probe.

*CBM math score* refers to the student's total number of digits correct on a sample of items, which represents the pool of problem types the student is expected to know by the end of the school year in a specific grade (Fuchs et al., 1990).

*Relative-age* refers to the month of birth within a specific calendar year in relation to school enrolment. In this study, there are three relative-age groups for students placed in the appropriate grade for their age. The three groups are defined below as: oldest, average and youngest.

*Youngest age group* (group 1) refers to students who were born during the months of October, November and December. They would be the youngest students in terms of years and months for any given grade level group (the school enrolment cut off date in British Columbia is December 31<sup>st</sup>) (Hedekar, 1997).

*Average age group* (group 2) refers to students who were born during the months of April, May, June, July, August and September (Hedekar, 1997, p. 24).

*Oldest age group* (group 3) refers to students who were born during the months of January, February, and March (Hedekar, 1997, p. 24).

*High performance group* (group 1) refers to students who achieved above the 75th percentile on the cumulative CD scores from the three CBM math norming periods.

*Average performance group* (group 2) refers to students who achieved from the 25th percentile to the 75th percentile on the cumulative CD scores from the three CBM math norming periods.

*Low performance group* (group 3) refers to students who achieved below the 25th percentile on the cumulative CD scores from the three CBM math norming periods.

*Retained students* refers to any student who was born in the calendar year previous to their present grade level peers. The reason for the student's grade retention is not known to the researcher and therefore could include students held back from beginning school or retained by parents or the education system for any number of reasons (Hedekar, 1997).

### *Research Questions*

1. Do markers produce consistent, reliable results (given that prior data analysis and presentation convey the sense that there was uniformity of the marking of the sample) using CBM math data as the measurement tool?
2. Are gender differences evident in all grades and at all norming periods as measured by CBM math data?
3. Are relative-age differences evident throughout all grades and norming periods when measured with CBM math data?

4. Are the CBM math results for gender and relative-age of the same magnitude and direction as the results produced by Hedekar & MacMillan's study (personal communication, January 30, 2002) for the CBM reading and writing fluency? *Writing fluency* for Hedekar and MacMillan refers to words spelled correctly (WSC).
5. Are the gender differences from this study, of the same magnitude as other research of gender differences when using CBM math data?
6. Are gender differences evident in different performance groups of all grades and norming periods as measured by CBM math data?
7. Are retained students equally, or more successful than students in their age appropriate grade as determined by their mean CBM math scores?

### *Hypotheses*

The following hypotheses are derived from the research questions and were tested during this study. Each hypothesis number reflects the number of the corresponding research question.

1. The first question regarding Inter-Rater reliability could not be formulated as a hypothesis. If markers marked the same, the means will be equal, variability will be zero and the correlation will be 1.00.
2. Investigation of the next research question requires that the mean of the math fluency, as measured by the correct digits on math probes, of the male students to be equal to the math fluency of female students, within each grade.

$$H_0: \mu_{gm} = \mu_{gf}.$$

$$H_1: \mu_{gm} \neq \mu_{gf}.$$

Where g refers to Grades 1 through 7, m and f refer to male and female.

3. For investigation of this question it was necessary to determine the means of three different age groups within each grade. These are the same relative-age group definitions used by Hedekar (1997).

$$H_0: \mu_{gj} = \mu_{gj'}$$

$$H_1: \mu_{gj} \neq \mu_{gj'}$$

Where j and j' = 1, 2, 3 for the three relative-age groups but  $j \neq j'$ . The other symbols are defined as in the previous questions.

4. This question regarding effect size differences between Hedekar & MacMillan's study (personal communication, January 30, 2002) and this present study, cannot be formulated as a hypothesis.
5. As with the previous question, effect size differences between other math research and this study, cannot be formulated as a hypothesis.
6. Investigation of this question required each gender from Grades 1 to 7 to be divided into three performance groups. The three performance groups were low, average and high groups (see definition section for further information).

$$H_0: \mu_{gmp} = \mu_{gfp}$$

$$H_1: \mu_{gmp} \neq \mu_{gfp}$$

Where p defines the performance groups. The other symbols are defined in question one.

7. To investigate the last question, the mean score of the students who were retained and are therefore older than they should be for their grade were compared with the scores



of students who are the correct age. Retained students, who are at least one year older than most peers in the same grade, should perform as an “average” student who is in the age appropriate grade.

$$H_0: \mu_{gj} = \mu_{gr}.$$

$$H_1: \mu_{gj} \neq \mu_{gr}.$$

Where r stands for the retained students. Other symbols are defined in the previous questions.

## CHAPTER TWO: LITERATURE REVIEW

There are three sections within this chapter reviewing literature relevant to this study. The first section investigates the development and use of CBM. Then, the second section examines research in mathematics achievement as it relates to gender differences, gender differences and the high performance group, relative-age differences, gender and relative-age differences using CBM math, and grade retention. Finally, section three reflects on the significance of the proposed study.

### Curriculum-Based Measurement

#### *Development and Use of Curriculum-Based Measurement*

Alternate forms of assessment and evaluation evolved as educators expressed dissatisfaction with traditional assessment practices (Daniels, 1999). The advantages of alternate forms of assessment over traditional, commercial, norm-referenced, and standardized tests are that teachers obtain a more accurate or representative description of a student's strengths, weaknesses, and needs. Using information obtained from alternate assessments allows teachers to develop individualized programs of instruction that improve the quality and effectiveness of their instruction. Curriculum-Based Assessment (CBA) is one form of alternate assessment that has been developed. Fuchs and Deno (1994), and Tucker (1985) state that three advantages accrue from the use of CBA: (a) it assesses the curriculum of the local school, (b) it provides local control over assessment, and (c) it allows teachers to assess a student's progress in relation to the curriculum. CBA data can assist in making decisions regarding the effectiveness of Individual Education Plans (IEPs) (King-Sears, Cummings & Hullihen, 1994). The five steps

involved in using CBA techniques in a classroom include: (a) analyze the curriculum, (b) prepare procedures and probes to meet the curriculum objectives, (c) probe frequently to collect the data, (d) display the data using a graph format, and (e) interpret the results-revisions and make decisions (King-Sears, Burgess, & Lawson, 1999; King-Sears, Cummings, et al.; Salvia & Hughes, 1990). As Marston and Magnusson (1988) state, CBM emerged as one type of CBA model. Curriculum-Based Measurement analyzes results from direct and repeated measurement procedures, administered for a specific length of time.

Curriculum-Based Measurement has become a familiar term to many educators since its development at the University of Minnesota by Stanley Deno and Phyllis Mirkin during the late 1970s and early 1980s (Deno, 1985; Marston, 1989; Shinn, Nolet & Knutson, 1990). The criteria set for the development of CBM measures requires them to be reliable, valid, simple, efficient, easily understood, and inexpensive (Deno, 1985). Fuchs and Fuchs (1997) state CBM was originally developed “to monitor student progress and ... link instructional planning with assessment information to enhance student outcomes” (p. 4). This includes making psycho-educational decisions, identifying students needing special service, and developing IEPs (Fuchs & Fuchs, 1992). These original purposes of CBM continue to motivate its use in many places, including within SD57.

The CBM tools developed to measure student progress during each testing time are called probes. Individual teachers can design CBM probes. It is possible to reference CBM measures to individuals, peers, or the curriculum. This allows for the development of local norms, whether peer, class, school or district norms; to facilitate decision making

(Shinn et al., 1990). Graphing student responses facilitates the CBM monitoring process. Graphs assist teachers in determining if a student achieved adequate progress or if the teacher's instructional plan requires modification (Allinder & Eccarius, 1999). Classroom and special education teachers use CBM measures to assist in determining the content a child needs to learn and the child's rate of learning (Howell et al., 1993). Curriculum-Based Measurement methods allow decisions to focus on the student's specific problem. Judgements regarding the student's problem can be made in a local context and can vary as the contexts change (Fuchs & Fuchs, 1997). Another advantage of CBM measures is their use in a problem-solving model to assist educators in making student performance and programming decisions (Deno, 1989).

*Limitations of Curriculum-Based Measurement.* Deno (1985) recognized that CBM has limitations. Developments in some instructional areas such as reading have progressed more than in other curriculum areas, such as math. Creators of CBM measures were able to agree on the "primary functional purpose of reading" (Deno, 1985, p. 230) compared to the function of other curriculum areas. The "lack of agreement [among the curriculum developers,] regarding the essential knowledge and skills to be required of all students" (Deno, 1985, p. 230) is hampering the development of CBM measures in other subject areas. While it is possible for educators to develop ad hoc CBM measures, they are limited in their use since their technical adequacy is unknown. Howell et al. (1993) noted that even when reaching agreement about the specific skills to test, there are limitations to the CBM tests due to their focus on basic skills. For example, in math, it is not possible to infer that a student's "skills in problem solving or mathematics application" (Howell et al., p. 169) are related to their scores in math computation.

Lombard (cited in Fuchs & Fuchs, 1997) rejected CBM due to its focus on academic behaviour while omitting intellectual assessment. Fuchs and Fuchs (1997) acknowledge that CBM is a static disability model, that defines what a student cannot do, and “focuses exclusively on the level of a student’s performance at one point in time” (p. 6). Although it is possible to use CBM to determine when to modify instruction for a student, Frank and Gerken (cited in Allinder, 2000) indicate CBM measures are unable to inform the teacher about what instructional techniques to implement. These are not the only concerns researchers have raised regarding the use and implementation of CBM, which educators must acknowledge when implementing CBM techniques.

Kranzler, Miller, and Jordan (1999), investigated racial/ethnic and gender bias on CBM reading as an indication of reading comprehension with African American and Caucasian male and female regular education students across grades 2-5. A biased test existed, according to Kranzler et al., if the regression lines of the groups significantly differed at either the intercepts or the slopes. Because they determined “CBM reading is not an unbiased test [as] the meaning of the scores on CBM differed across race/ethnicity and gender at particular grade levels” Kranzler et al., raised concerns regarding the use of CBM for screening, determining special education eligibility, and termination of services. Therefore, CBM scores did not mean the same for each subgroup of the sample population. According to Kranzler et al., the greatest concern regarding biased estimation of CBM scores is for students near the cutting score, which determines eligibility for special education and related services. However, Kranzler et al. did not determine why the results of the study were not consistent across the grades. In fact, for Grades 2 and 3 no evidence of test bias was apparent. Kranzler et al. also did not attempt to acknowledge

that differences between subgroups found in the study might reflect real performance differences between the groups. Nor do they indicate if the creators of the CBM reading measures attempted to reduce sources of test bias. Instead, the assumption is that group differences in average scores imply a test is biased. The inference is that one group is less able than the other (Gipps & Murphy, 1994).

As found by Kranzler et al. (1999), if CBM determines that differences exist between subgroups of a student population, then it is important to recognize these differences and the implications they might have on student achievement. Knowing if differences exist would be extremely beneficial to school districts. Knowledge that differences exist in subgroups could indicate that using the same CBM norming score for each student in a grade may not be realistic. Therefore, comparison to students in another subgroup may not provide an accurate representation of the student's performance. These concerns indicate the need to determine the existence of gender and relative-age differences in the CBM math data.

*Inter-rater reliability of Curriculum-Based Measurement.* Given that development of CBM measures enabled a large number of educators to administer and mark the probes, one initial concern regarding CBM measures is marking reliability. What adds to the usability of these measures is interscorer agreement (Marston & Deno cited in Fuchs & Fuchs, 1992), which allows student comparison even when several different people administer and score the probes. Inter-rater reliability (also referred to as interscorer agreement or interjudge reliability) is the extent to which raters or markers agree on the score or the reported data (Sax & Newton, 1997). Marston (1989) states that reading measures have interscorer agreement coefficients of .99. In spelling, Marston

reports interscorer reliability coefficients for words and correct letter sequences of .99 and .91 respectively. A summary of math measures by Fuchs, Fuchs and Hamlett cited in Marston, indicated the math interscorer agreement on a sample of 30% of the protocols is .98. Interscorer agreement on CBM math measures, determined by Tindal, Marston et al., cited in Marston, produced a range of reliability scores from .90 to .99. Allinder's (2000) study combining teachers' self-monitoring of instructional strategies with CBM in math computation determined the interscorer agreement, calculated on 15% of the tests, was 99%. A study by Allinder and Eccarius (1999) researched CBM reading procedures for deaf and hard of hearing students who used manually coded English. Interjudge reliability between two judges for passage reading ranged from 40% to 100%, with a mean of 78.69, standard deviation of 13.59 and a median of 81%. For the second aspect of the study, Allinder and Eccarius computed the interjudge reliability in writing. Allinder and Eccarius independently compared the student's story retelling with idea units. On 50% of the retells, the interjudge reliability mean was 78.76, the standard deviation was 8.75, and the median was 100% with a range of zero percent to 100%. Another CBM study of written expression proficiency of middle school students by Espin et al. (2000) required two raters to mark 20 randomly selected protocols. The percentage of agreement between the markers in the Espin et al.'s study "was calculated by dividing agreements by agreements plus disagreements and multiplying by 100". Interscorer agreement of the writing samples ranged from 85.63% to 89.66% in Espin et al.'s study. Most studies report high positive values for average inter-rater agreement or reliability. Correlations of less than 1.00 indicated some rater disagreement existed. Reported ranges of inter-rater scores add a further look at rater disagreement. An example is

Allinder and Eccarius' study, which reported a range of agreement between the markers of zero to 100%, despite a mean was 78.76, and a high median value of 100% for story retell with idea units. This study by Allinder and Eccarius provides evidence that rater disagreement can exist. What is unknown, at this time, is if the results reflected in these studies reflect the results of the SD57 CBM math probes.

One study provides evidence regarding rater reliability of CBM reading and writing measures by the markers of SD57. Hedekar and MacMillan (personal communication, January 24, 2002) confirmed a high degree of Inter-rater reliability between 10 randomly chosen markers. For Hedekar and MacMillan's study the median inter-rater reliability correlation for total words written (TWW) and writing (WSC) was a median of .99 with a range of .97 to 1.00, and for words read correctly the median was 1.00 with a range of .99 to 1.00. All identified literature, including the SD57 research, reports that inter-rater reliability for CBM measures is high when reported as a correlation or the median of interscorer agreement. If a high degree of marking consistency among markers of student CBM probes exists, as indicated by Hedekar and MacMillan, then educators gain confidence that results among markers are consistent, and therefore comparable.

What presently is unknown in SD57 is the degree of consistency among the markers of the CBM math norming probes. Despite present use of the CBM math probes and norming data, to date no data collection has determined the level of inter-rater reliability in CBM math for SD57. Until this happens, the usability and comparability of the CBM math measures administered by the different markers in SD57 is suspect.



*Curriculum-Based Measurement Research in School District No. 57*

In 1997, the first CBM research project took place in SD57 independent of the norming technical report (School District No. 57, 1996). Hedekar (1997) undertook a study using the CBM reading and written expression data collected for the development of the first SD57 CBM norming project. This study focussed on gender and relative-age differences in reading, and writing fluency. Writing fluency--also referred to as written expression--was measured by the number of total words written (TWW) and the number of words spelled correctly (WSC). According to Hedekar, "a consistent gender effect was found at all grade levels. Male students' mean score in reading, writing and spelling was lower than female students' mean score at every grade level" (p. ii). Hedekar did not find a relative-age effect for reading and writing fluency at any grade.

Further analysis of Hedekar's (1997) results is presently underway. Personal communication with Hedekar and MacMillan (January 17, 2001) indicates they are currently repeating the analysis of the first CBM reading and writing study using a doubly-multivariate design. Initial indications are that the results are similar to those obtained by Hedekar in the original analysis.

After identifying the need for a standardized, norm-referenced assessment tool in mathematics at the elementary level, the second major SD57 CBM research project took place during the 1999-2000 school year. This study followed procedures established in developing the initial reading and writing norms. Once again, 20% of the elementary students from Grades 1 to 7 participated in the study. As in the first CBM norming project, collection of data took place during the three norming periods of October, January and April for Grades 2 to 7. Grade 1 students only participated during the April

norming period. Walraven and MacMillan (2000) presented the results of this CBM math norming project in the Draft Technical Report of the CBM (Math) Norming Project (see Appendix B). Educators received the CBM Math Draft Norms Tables at an inservice held in September 2000 (School District No. 57). Although this study developed norms for the CBM math data collected, Walraven and MacMillan did not determine if gender, relative-age, performance group, or grade retention differences existed in the data. Because Hedekar (1997) found evidence of gender differences, but not relative-age differences for the CBM reading and written expression data, the present study of gender and relative-age in the CBM math data is necessary. In addition, Walraven and MacMillan did not investigate the inter-rater reliability of the markers of the CBM math. On September 18, 2000, Walraven commented to this researcher that some of the trained markers appeared more severe than others were on certain marking criteria including alignment of digits. Walraven's information, along with the questions and concerns raised by some teachers attending the inservice, confirm the necessity of investigating the reliability of the CBM math markers.

In 2000, a validation study of the original CBM reading and written expression study in SD57 took place. Fewster (2000) wanted to "examine the validity of CBM scores for predicting later junior secondary school achievement, [and] to verify its adequacy as a standardized indicator of student performance" (p. 4). Using the CBM reading and written expression scores from the initial CBM project, Fewster compared the CBM scores for 678 Grade 6 and 7 students to their year-end marks in English and Social Studies for Grade 8, 9, and 10. She concluded CBM scores are useful "as indicators of student performance in the basic skills of reading and written expression"

(p. 80) and provide “information when making decisions about students based on their academic performance” (p. 81). Fewster’s study determined it is possible to use CBM scores to differentiate between student performance groups. CBM can identify students requiring remedial support. In addition, CBM reliably separates students entering honours programs from those entering regular classes and separates students in regular classes from those requiring Special Education and Learning Assistant services. When allocating further resources for students when using CBM as a screening measure this ability to differentiate between performance groups at both ends of the spectrum provides confidence for SD57 personnel. While Fewster determined that CBM reading and written expression measures have predictive validity, it is important to know prior to using it for future predictive research or comparison of results from different measures if subgroup differences exist within the CBM math data.

The research in CBM demonstrates its feasibility as an alternative to standardized testing. CBM has proven to be a reliable tool that can establish differences between groups of students in reading and writing fluency. Research in SD57 demonstrated CBM reading and writing fluency measures are good predictors of future marks in Humanities subjects. Whether CBM math can ascertain differences between groups of students as did the CBM reading and writing is unknown. Successful completion of this present study provided confidence regarding the reliability of the CBM math measures.

This present research project stems from the CBM math norming project conducted in SD57 during the 1999-2000 school year. Data for the CBM math norming project were already collected and available for this researcher to analyze. The purpose of my research was to investigate if gender, relative-age, performance group and grade

retention differences can be determined within the CBM math norming data. Through this study, further investigations into gender and relative-age differences using CBM measures are possible. My study will provide educators with the opportunity to examine these differences in the area of CBM math and enable them to be compared to the results from Hedekar's and MacMillan's (personal communication, January 30, 2002) study of CBM reading and written expression.

### Research in Mathematics Achievement

There are three main topics are of interest within the area of math achievement. The first topic of interest is gender differences in math. The second topic is relative-age and grade retention is the third subject of interest.

#### *Gender Differences in Math*

The first issue worth consideration is the purported math gender gap between males and females. Historically, boys have performed better than have girls in the areas of math and science. Beal (1999), summarized that differences in math-fact retrieval between males and females in mathematics performance lead to males outperforming females in testing situations.

Sadker (1999), indicated that the gap between boys and girls in math has been declining. Cole, (1997) examined data for Grades 4 to 12 derived from several national studies in the United States and from a few international studies. From her research, Cole found that the math gender gap favouring males is significantly smaller than 30 years ago. Cole's study measured gender differences using the standard mean difference ( $D$ ). Calculation of the  $D$  is the same as Cohen's  $d$  (Cohen, 1992). The Grade 4 standard mean

difference was trivial in magnitude with a value of less than 0.1 in favour of girls. By Grade 8 the boys were favoured also by less than 0.1 (Cole, p.15). Sadker and Cole both agree that the gender gap in math is decreasing. The question arises as to whether the purported decreasing math gender gap during recent years reflects results found by other researchers.

Cole (1997) however, also discovered that while gender differences in math concepts at Grade 4 are small, males increase their advantage significantly from Grades 8 to 12. Recently completed research by Leahey and Guo (2001) used two large representative longitudinal math surveys. Leahey and Guo also concluded that males and females begin with equal starting points in elementary school but boys demonstrate a “faster rate of acceleration.” This results in a “slight, late-emerging male advantage in the general population” (Leahey & Guo), which went unchallenged when limited to high-scoring students. It is possible this trend of an increasing male advantage in math until Grade 12 is reflected by students throughout elementary school.

The American Association of University Women Educational Foundation, AAUW, (1992) investigated the educational experience of girls in the United States. According to the AAUW, girls are not receiving the same educational experience as boys. However, existing gender differences in math favouring boys, measured by recent meta-analyses, are very small and on the decline. These results showed that at age nine no evidence of gender differences existed; by age thirteen minimal differences existed and by age seventeen larger differences existed which favoured males. According to the AAUW the age of the sample, the cognitive level of the test and the academic selectivity of the test influence the existence of gender differences. These results indicate that gender

differences are nonexistent in younger grades but increase as students move through higher grades.

A reanalysis of data from an international study conducted in 1991 confirms that gender differences in math are not limited to one location (Beller & Gafni, 1996). The data from the International Assessment of Educational Progress (IAEP) included representative samples of approximately 3,300 students from each of seven countries, including the United States but not including Canada. The mean total score between nine-year-old boys and girls were not significant. However, nine-year-old boys were significantly different from girls in the measurement subdomain. In contrast, all results for thirteen-year-olds were significantly different and favoured boys, with the exception of the algebra subdomain. With one exception, the effect size results for all subdomains and score totals were trivial but consistently in favour of boys. Calculation of effect sizes used the mean performance for boys minus the mean for girls, divided by the standard deviation, computed across the two groups. Total score effect sizes were 0.04 and 0.12 for nine and thirteen-year-olds respectively (p. 369). According to effect size definitions by Cohen (1992), these effect sizes are trivial. A further study of thirteen-year-olds by Beller and Gafni (2000) also used IAEP data for 1991 and 1988 found effect sizes of 0.11 and 0.03 respectively in favour of boys. Again, the effect sizes were trivial. Although these studies verify the existence of gender differences that favour boys, the results are not consistent across ages and in effect size.

Earlier Canadian research in Manitoba by Morrow and Goertzen (1986) determined that when gender achievement differences existed they were “usually small and favour[ed] males” (p. 5). These Canadian results are not the only ones indicating

gender results in Canadian schools do exist in some circumstances. Over a decade later in June 1997, the International Association for the Evaluation of Educational Achievement (IEA) reported on Mathematics Achievement in the Primary School Years in the Third International Mathematics and Science Study (TIMSS). Statistically significant gender differences at  $p < .05$  in the TIMSS study at the third grade found boys scored higher than did the girls in the content areas of measurement, whole numbers, and mathematics overall (Mullis, Martin, Beaton, et al., 1997). Trend comparison of gender differences from the 1995 test to the 1999 test in the IEA TIMSS 1999 study for the eighth grade found no significant changes. In Canada, where gender differences in math exist, they usually favour boys.

Gender research indicates the differences appear to favour boys. The evidence confirms there are reasons to be concerned about the existence of gender differences. However, research has not found the differences to be consistent across all grades or ages.

*Gender differences favouring girls.* While the majority of research results on gender differences favoured boys over girls, the results from Hay, Ashman and van Kraayenoord (1998) contradicted other studies. In an Australian study of 390 Grade 6 elementary students from 18 schools, they found that girls outperformed boys in mathematics, reading, and spelling. A new question to ask is if girls are outperforming boys in math. The IEA TIMSS November 1996 report for the Middle School Years established that in Grade 8 the gender differences favoured the girls (Beaton et al., 1996). These results change the concerns from boys outperforming girls in math, to girls outperforming boys.

*No evidence of gender differences.* Gender differences are not apparent in all math studies undertaken. The IEA TIMSS 1997 study demonstrated in the fourth grade that statistically significant gender differences in math did not exist (Mullis, Martin, Beaton, et al., 1997). According to the IEA TIMSS November, 1996 report for the Middle School Years gender differences were not significant for seventh and eighth grades (Beaton et al., 1996). The TIMSS 1999 International Mathematics Report on the eighth grade indicates a lack of significant gender differences in specific math content areas (Mullis, Martin, Gonzalez, et al., 1999). According to these studies, gender differences in math achievement do not exist in all situations.

In an attempt to refute the AAUW report, Kleinfeld (1998a) reviewed information on several measures to demonstrate that schools are not shortchanging girls. Her results showed that the standard mean difference for Grade 12 students using national samples of students, for math computation and math concepts are .18 and -.11 respectively. In math computation, girls were favoured but in math concepts, boys were favoured. According to Kleinfeld neither difference is large enough to be considered a small effect and therefore negligible gender differences are evident in the general population.

A study by Willingham and Cole (1997) used tests administered to national samples of Grade 12 students and determined no gender differences exist in several subject areas. The average standard mean difference of 14 math tests was -0.11 in favour of boys (p. 59). In fact, all but two math tests show a small or negligible standard mean difference in favour of boys. The other two math tests demonstrate a trivial standard mean difference that favours girls. Willingham and Cole acknowledge most of the differences did not reach the “small” level as determined by Cohen (cited in Willingham



& Cole). The average math standard mean difference cannot even be considered “small.” Therefore, gender differences in math achievement were trivial.

Hall et al. (1999) tried to demonstrate gender and racial differences in Grade 5 and 8 students. Their sample had 74 participants, of which 36 were girls. The researchers accessed the student’s California Achievement Test (CAT) scores. Scores from the math calculation and math concepts sections of the CAT were the dependent variables. Gender and race were the independent variables. The data was analyzed using a MANOVA. While the sample size appears small for this analysis, Mardia (cited in Tabachnick & Fidell, 2001, p. 329) assures robustness with 20 cases in the smallest cell. From the analysis, Hall et al. discovered “significant differences for race but not for gender” (p. 5). However, they do acknowledge significant differences may not occur until students reach higher grades. Although Hall et al. did not find evidence of gender differences in math, they did not reach the conclusion that differences no longer exists in high school.

Ma (1999) investigated gender differences in achievement at the end of Grade 7 and rate of growth from Grades 7 to 11. The study analyzed a stratified American national sample with 3,116 students from 52 schools using a three-level hierarchical linear model (HLM). Ma’s results showed “there were no significant gender differences in either the grade 7 status or the rate of growth” (p. 457). The lack of significant difference in growth from Grade 7 to 11 contradicts results that found boys had an advantage over girls in their rate of growth in math skills (Leahey and Guo, 2001).

Present research into gender differences in math shows a strong possibility that no sizable differences exist. Although the differences are small, there remains a lack of agreement among researchers regarding their existence. This lack of agreement for the

general population remains an ongoing issue. It is necessary to investigate this issue in order to address any differences that arise from the data. Therefore, due to the debate regarding gender differences, it is impossible to generalize the findings. Analysis of the CBM math data is necessary to determine if elementary school gender differences in math computation exist or are of a magnitude that causes educators concern.

### *Gender Differences in the High Performance Group*

If the gender gap in math achievement is decreasing, why then is there cause to be concerned? Kleinfeld (1998b) answers that there is “greater male variability” (p. 49) than female variability in the population. Willingham and Cole (1997), and Cole (1997) analysed national test batteries from Grades 4 to 12 and confirmed that a pattern exists showing gradually increasing greater male variability than female variability. Research on math-fact retrieval by Royer, Tronsky, Chan, Jackson and Marchant (1999) verified the existence of more male variability than female variability. More male variability should be evident in a greater range of scores for males than females. It is possible that the existence of greater male variability in learning is a result of a distribution that is bimodal, trimodal, or heavily skewed rather than a normal distribution. Kleinfeld demonstrated the existence of gender differences in learning disabilities, with information showing the number of males to females with learning disabilities is three to four times higher. Thus, boys arrive at the bottom of the ability group. Conversely, “even if the difference is small in the population as a whole, far more males will show up in the visible category of top performers” (Kleinfeld, 1998b, p. 62). Kleinfeld (1998a) used the top 10% of the students to define the top performers and found males outperform females in math. Research by Willingham and Cole (1997) found a standard mean difference of

- 0.15 favoured boys for students taking Advanced Placement Tests. While the information from these studies is based on American research, it would be valuable to determine if these issues exist in other locations including Canada.

Willingham and Cole (1997) noted another issue exists for students who score in the top end of math achievement. They noticed gender differences favouring males is increasingly apparent when students write advanced tests in math. Within the top 10% of the Grade 12 students, Willingham and Cole, found a female to male ratio of .7. Even at Grade 4, Willingham and Cole found there were more males than females among the top 10% of the students. Royer et al. (1999) concluded that on math achievement boys perform better than do girls. In fact, Royer et al. also found fast males perform faster than fast females. Research by Beller and Gafni (2000) suggests “boys do relatively better than girls as items increase in difficulty” and boys answer more of the difficult questions than girls do.

A study from the U.S. by Fan (1995) used a national sample from a longitudinal database, which tracked approximately 25,000 students in eighth grade. Item Response Theory (IRT) was used to equate the difficulty level of the multiple test forms administered. Fan demonstrated that gender differences in math do not appear to exist when comparisons take place using measures of central tendency and any effect sizes in most cases would be small. However, when comparisons focus on high achievement, a meaningful gender difference is evident. Gender differences in the high performance ranges increase as students proceed from Grade 8 to 12, and as the comparison progresses to the highest proficiency level. In Grade 8, in the first quartile, 51.14% are males; however, in Grade 12 at the 95th percentile male students outnumber female students by

a ratio of 2:1. It is presently unknown how these results generalize to other measures of math achievement.

Hedges and Nowell (1995) analyzed six American national samples and investigated the issues of gender as they related to variability in scores, ratios of high achieving students and mean differences. Their findings confirmed that in math “the variance of male scores is larger than that of female scores” (p. 44). The variance in male and female scores had changed very little over time. The study also showed that males performed better than females in math. Using the data from the National Educational Longitudinal Study of the Eighth Grade Class of 1988, the standard mean difference for mathematics was 0.03 in favour of boys. A nonsignificant difference in variance (1.06) also favoured boys. Difference in variance was calculated as “ratios of male score variance to female score variance” (p. 43). The ratio of the number of males to the number of females for students in the top five percent of the national distribution was 1.64 with a standard error of 0.18 for the same mathematical data, in favour of boys. As Hedges and Nowell conclude, small differences in means combined with variation differences can influence the number of students excelling in careers requiring these skills. Thus, it is important to determine whether these results are a realistic reflection of all math achievement.

In contrast to other researchers of high performing students, Mullis, Martin, Fierros, Goldberg, and Stemler (2000) found no statistically significant difference at the .05 level in gender in the top 25% of Canadian Grade 8 students who participated in the third TIMSS study. This also held true for the number of males compared to females who were in the top 25% of the Canadian Grade 4 students in the Third TIMSS study.

Results from the TIMSS study indicated gender differences are not always evident amongst the high performing students.

If differences do exist between the boys and girls at the top performance groups, then it is important to determine and acknowledge these differences. However, the research regarding differences amongst the high performing students is not consistent. While most researchers investigating gender differences for high achievers agree differences exist, controversy surrounds this question. Action to counteract negative impacts on high performing students' math achievement differences cannot be undertaken without first establishing it exists in all math achievement. As Fan (1995) concluded, educators must understand the students who score in the high ranges of math achievement will "likely become our future scientists, engineers, chemists, [and] physicists" (p. 16). Therefore, educators cannot afford to be complacent because an apparent lack of gender differences exists. It is imperative to confirm potential gender differences between high achieving students in math before educators can address any differences that may exist.

The concerns raised by Kleinfeld (1998b), Willingham and Cole (1997), Cole (1997), Fan (1995) and Hedges and Nowell (1995) regarding gender differences and ability are not limited to the U.S.A. In the Executive Summary of the TIMSS 1999 study, gender differences among high-performing students is described as significant even though the actual difference may be small (Mullis, Martin, Gonzalez, et al., 1999). Gender differences within different ability groups are a concern. Therefore, it is necessary to establish their consistent existence in math achievement.

### *Relative-age Differences*

Parents and educators often question if younger students will experience the same academic success achieved by older students in the same grade. *Relative-age* refers to the month of birth within a specific calendar year in relation to school enrolment. Three categories or groups pertain to relative-age in this present study. Recall that the youngest group consists of students born from October to December. The oldest group encompasses students born from January to March, and the average group includes students from the months of April to September. Relative-age is a concern if younger students, born during the last few months of a calendar year, do not perform as well as their older peers who were born earlier during the same year. Relative-age considerations are important if the older students are more successful academically. However, not all studies calculate relative-age in the same manner (Boyd, 1989; Olson, 1989; Rabinowitz, 1989; Warder, 1999).

Research into relative-age differences suggests that relative-age influences achievement. Warder (1999) examined literacy skills in Kindergarten, first and second grade students from a total of six classes. The students were divided into three relative-age groups each defined by a third of the year. She found achievement decreased with younger students when comparing the percentage of students achieving specific literacy skills at grade level. Warder did not use a statistical analysis to determine if the differences she observed were significant. Therefore, it is impossible to determine if her results reflected a real difference in her sample, which can be inferred to other students.

Rabinowitz (1989) determined entry age (relative-age) was a significant factor on the scores of 83 Grade 1 to 6 students. Students in the *early* entry group had their sixth

birthday after August first the year they started first grade. *Middle* entry students turned six between January first and August first the year they began first grade. The *late* entry group included students who turned seven before December 31 the year they started first grade. Using scores obtained on the Iowa Mathematics Achievement Test, Rabinowitz found relative-age impacted math achievement. An ANOVA determined a significant difference at  $p < 0.1$ . Rabinowitz set this significant level as a cautious approach for first grade placement decisions. The actual value of  $p$  was .07. Typically, a  $p$  value is set at .05, or even .01 for a cautious approach to reduce Type I error. Therefore, the results of this study do not demonstrate a significant difference between relative-age and math achievement as Rabinowitz suggests.

Olson (1989) determined if relative-age has an effect on elementary school performance. His study followed 6,246 students for six years. Each year the Iowa Tests of Basic Skills were administered. Four relative-age groups were identified. Each group encompassed students born within the three months defining the group. A MANOVA showed a significant difference for reading and math at  $p < .01$  for both subjects. Further analysis did not find differential rates of achievement growth in either subject at  $p > .1$ . Olson found a consistent performance differential in mathematics maintained by younger students of approximately three tenths of a year behind older student, throughout their elementary school career. Thus, according to Olson, relative-age is a concern. While these studies allude to potential relative-age differences, further investigation is required before concluding they consistently exist.

In another study Narahara (1998b) reviewed research regarding school entrance age and academic advantage for older children over younger ones in the same grade and

found that “the research findings often contradict[ed] each other” (p. 15). Only one study by Cameron and Wilson (cited by Narahara) used math scores as well as reading scores to calculate relative-age differences. Other research reviewed used reading not math. The studies examined showed an advantage for early grades that in later years was nonexistent. Smith and Shepard (cited in Meisels, 1992) found similar results. Because the studies defined relative-age in a variety of ways it is impossible to conclude if relative-age differences consistently affect the achievement of all students.

Bisanz, Morrison, and Dunn (1995) investigated the effects of age and schooling on conservation of number and mental addition. A cutoff design analyzed three groups of Kindergarten and Grade 1 students whose birthdays were two months before or after the March school entry cutoff. Altogether 56 students participated. Data analysis used a  $3 \times 2 \times 2$  ANOVA with repeated measures. A significant difference between the groups was found at  $p = .012$  for conservation of number. Mental arithmetic was significantly impacted by length of school experience at  $p = .001$ . Conservation of number increased with age but accuracy of mental arithmetic improved with an increase in schooling. Although relative-age influences some math skills, it is not responsible for all math achievement.

Gullo and Burton (1992) studied age of entry and sex as factors in academic readiness for kindergarten. In contrast to other studies, Gullo and Burton found age of entry was one of the three factors contributing to “prediction of academic readiness at the end of kindergarten” (p. 183). Sex as a variable did not significantly account for academic readiness in this study. The 4,539 students took the Metropolitan Readiness Test, Level II, Form P (1974) in May of their five-year-old Kindergarten year. First, a



hierarchical regression analysis explored the effects of children's age, length of preschool experience, and gender on academic readiness. Administration of the Cooperative Play Inventory-Revised (CPI-R) assessment and screening instrument controlled for students "at-risk." An ANCOVA using the CPI-R score as a covariate found main effects for age at  $p < .001$  and length of preschool experience at  $p < .001$ . The results also found a significant interaction for age and preschool experience at  $p < .001$ . From this study, it is apparent that school entry age is a contributing factor to academic success.

*Relative-age differences in the high performance group.* Relative-age could be one factor impacting the achievement of students in the high performance group. Sweeney (1995) undertook to determine the age children should begin attending school in this group. High ability was determined as achieving 129 or better on the Cognitive Abilities Test. The 275 students from Grades 2 to 8 were divided into three sections by birthdate. Students in the second trisection were excluded to provide more contrast between the first and last trisection groups. Results of the three-way ANOVA produced a significant main effect ( $p < .05$ ) for age position as well as grade and gender. Further investigation will determine whether this significant difference for relative-age is critical to the academic achievement of other high ability students. If relative-age plays a vital role in the academic achievement of high ability or performing students, it is imperative to determine if differences do exist. Sweeney's research raised concerns that even if relative-age differences are not apparent using central tendency measures in the general population the differences might still exist within the top five percent of the population. It is therefore necessary to investigate possible significant relative-age differences in the math achievement of high performers, as generalization of Sweeney's results is not yet

determined. Further research is necessary to determine the impact of relative-age differences on high performance students.

*Relative-age differences and science.* Educators and parents are often anxious about the effect relative-age has on the success of young students in language arts and math. Expanding relative-age research to other subjects, grades, and ages could expand the concerns regarding the impact of relative-age on achievement. Bell and Daniels (1990) investigated the relative-age or birthdate effect on science achievement of eleven, thirteen and fifteen-year-old students. This British study used science data collected for four years with 12,000 to 15,000 students per year to determine if “the birthdate effect [relative-age] persists beyond the primary years.” Relative-age of each student was calculated in days. Bell and Daniels concluded there is a birthdate effect, which influences the academic performance of a student who is one of the youngest in a grade. If science and math are linked, then these relative-age differences may generalize to math research as well as science. Hence, if students’ are to achieve success in all academic subjects and grades relative-age differences are an important consideration to investigate.

*Relative-age differences and sports.* While controversy surrounds the impact of relative-age and academic achievement, research has also investigated the effect of relative-age on sports achievement. Whether linked to academic achievement or not, relative-age does influence sports achievement. Glamser and Marciani (1992) discovered relative-age plays an important role in major college football participation. Boucher and Mutimer (1994) concluded that professional hockey players’ benefit from a relative-age advantage. A study by Barnsley, Thompson and Barnsley (cited by Barnsley, 1988) concurred with Boucher and Mutimer’s findings that older players have the relative-age

advantage. When debating the influence of relative-age on academic achievement one must not overlook its impact on achievement in sports up to adulthood. If relative-age can play an important role in one aspect of a students' development, educators and parents cannot be complacent that its influence is limited to a specific aspect of development.

*No evidence of relative-age differences.* Other research concluded that the relative-age of a student does not influence academic achievement. Gredler (1992) reviewed literature regarding the influence of entrance age on student achievement. Several of the studies investigated math achievement as well as other achievement. From these studies, he concluded that younger-aged children at the end of Grade 1 and 2 obtained lower placement scores than other children, but the scores they obtained exceeded the placement score expected for that grade. He also noted younger children had a failure rate similar to other children. In fact, one study by Carrington (cited in Gredler) found younger-aged students achieved academically as well as older-aged students. According to Gredler, entrance age is not a factor impacting school achievement.

A study by Narahara (1998a), looked at the effects of school entry age and gender on both reading and math achievement in Grade 2. Her American study contained 24 Grade 2 students divided into three age groups, with each group comprising an equal third portion of the twelve-month age range. Using a standardized achievement test (the TerraNova) Narahara sought to determine if there was a correlation between performance in math and reading and the age at which a child enters kindergarten. She also looked for gender differences in reading and math performance of second grade students, but did not report if the differences were significant. From her study, Narahara found "there is a low

or negligible [nonsignificant] correlation [of .28] between kindergarten entry age and academic achievement” (p. 7). A significant correlation requires a value of .42 or greater. While Narahara provided evidence to support the hypothesis that relative-age did not influence academic achievement in math her study uses a very small sample. Therefore, the results might be an artifact of small sample size. However, according to Narahara, relative-age is not a concern.

Bickel, Zigmond and Strayhorn (1991) established that relative-age had a significant impact on math achievement when students entered first grade in a U.S. school district. Entrance age was considered a continuous variable for the 222 participants in the study. Bickel et al. investigated four outcome variables, two of which included math achievement. The major analysis included a covariate statistically controlled by partialling and computed partial correlations of the outcome variables with entrance age. Although Bickel et al. noted relative-age had an impact in first grade, four years later in Grade 5 there was no relative-age effect.

Boyd (1989) determined if differences existed in reading and math achievement between younger and older students in Grades 1 to 5. Two relative-age groups were created in Boyd’s study: Younger students entered Kindergarten at age five to 5.5, whereas older students were 5.6 years and older. Reading and math CAT scores were used for Grades 1 to 3. For Grades 4 and 5 reading and math scores from the Stanford Achievement test were provided. A repeated measure design determined no significant differences at the .05 level existed between younger and older students in any grade, in reading or math achievement scores. A MANOVA was used to investigate relative-age differences in achievement at the .05 level with other variables separate and in

combination including race, gender, and family income. Boyd did not find significant difference between younger and older students at any grade level for reading and math achievement. As a result, Boyd did not find evidence to suggest that relative-age influences math achievement.

Two studies using data from SD57 also investigated the effect of relative-age on students reading progress in Grades 2 to 7. Using a Many-Faceted Rasch model MacMillan (2000) compared the growth of reading fluency scores simultaneously with relative-age, gender, and reading probe difficulty. The Many-Faceted Rasch method of measurement provides a researcher with a method of obtaining objective, fundamental measures from several random variables of ordered category responses and then evaluates on a logit scale the responses of a set of persons to a set of items (Linacre & Wright, 1996). MacMillan concluded there was “a lack of effect due to relative-age” (p. 406) since the grouping order of oldest, average and youngest students did not remain consistent across the grades. If relative-age impacted achievement, then the same relative-age group order would consistently be achieved from grade to grade. In MacMillan’s analysis, this was not the case. In addition, the differences of mean ages of nine months (oldest-youngest) differences were not represented by an equivalent difference in reading fluency. Instead, only differences of the equivalent unit of one to two months reading fluency were apparent. Hedekar (1997) originally investigated relative-age differences using 3 x 2 (birth group by gender) between groups ANOVA. Her results determined that there was no advantage to relative-age on CBM reading scores. These studies indicate relative age does not affect achievement.

The relative-age research indicates that an ongoing debate persists regarding the impact of relative-age on academic achievement. No conclusive evidence exists to bring the relative-age debate to fruition. Even if differences do not exist within the general population a new concern that requires further research has emerged regarding the existence of relative-age differences within the high performing students. While disagreement exists between researchers regarding the effect of relative-age, there is no researched understanding regarding the effect of relative-age on the academic success of students using CBM math data. Therefore, it is vital to determine if relative-age is contributing to the academic success of students as measured by the CBM math.

#### *Gender and Relative-Age Differences Using Curriculum-Based Measurement Math*

It is difficult to know what results to expect regarding the outcome of math achievement when analyzing CBM math data by means of a conventional analysis. However, it was possible to obtain a glimpse at what the results might be. MacMillan (2001) performed a Many-Faceted Rasch measurement analysis of the CBM math data from 1477 Grades 2 to 7 students in SD57. MacMillan concluded that no significant gender differences existed from one grade to another. He also found significant relative-age differences existed for only two of the grades and therefore consistent differences were not evident. Another argument suggesting a lack of relative-age differences was the 2.1 month difference within a six month age span (p. 20). Despite the controversy surrounding gender and relative-age differences, MacMillan's study suggests no evidence of gender or relative-age differences exists in the CBM math data. However, personal communication with MacMillan (January 13, 2002) revealed that with further analysis a gender difference might exist. Therefore, until completion of further analysis on the

CBM math data uses conventional analyses, confirmation of MacMillan's results are not possible. What MacMillan's study does not answer is whether there are gender differences in the high performing students in the CBM math data. Even with MacMillan's results, this research on gender and relative-age differences remains vital to reach an understanding regarding the CBM math results.

### *Grade Retention*

Parents and educators often consider grade retention as a means to helping a student catch-up with their academic skills or to improve their marks. Retaining students in a grade appears to be a strategy considered for at-risk students who may have difficulty achieving passing marks.

As a technique to increase school performance by providing students with more time to develop skills, grade retention has received negative publicity. After reviewing articles on grade retention, Foster (1993) determined students did not benefit academically from grade retention. Owings and Magliaro (1998) concluded that grade retention has a history of failure and harms learners. Meisels (1992) also found "that [retention] produce[s] more negative effects than positive outcomes" (p. 171). In fact, Reynolds (cited in Owings & Magliaro) suggested that grade retention may be decreasing achievement, particularly in reading. This is confirmed by Meisels (1992) analysis of the National Educational Longitudinal Study (NELS), 1988. The NELS study also found students who had not been retained not only performed better on reading but also on math and science (Meisels, 1992). Students who were not retained demonstrated higher test scores and grades (Meisels & Liaw, 1993). Crosser (1991) compared the academic achievement of a group of students who entered school at age five with a matched group

that entered at age six. Crosser concluded that six years after the students started school at age six they did not differ in their achievement significantly from those who started at age five.

Shepard and Smith (1987) completed a study on Grade 1 students to assess the impact of Kindergarten retention on 40 students who had been retained. The researchers matched the retained group with a control group of 40 students who had not been retained. After analyzing both groups of students on several outcome measures, no differences were evident between the two groups of students, with the exception of results from a reading test. In math, the scores of both groups matched. This research refuted the belief that at-risk students benefit academically from an extra year in school.

All the studies on grade retention agree that retained students are not more successful academically than their age appropriate peers. The focus of all these analyses, however, centred on American research and did not indicate how grade retention in Canadian schools affects learners. If students are to successfully achieve academically, discerning the impact of retention on learners is important for educators and parents.

### Significance of the Proposed Study

It is important that this study be undertaken to establish the reliability of the SD57 markers who undertook the marking of the CBM math data. Much of the research indicates good marking reliability for CBM measures. However, Allinder and Eccarius (1999) reported a large range in their marker agreement, even when obtaining a high mean and median. If the CBM math data for SD57 has a large range of scores obtained by different markers, then it can be argued that not all markers reached agreement. Even



a high correlation as reported by Hedekar (1997) indicated some variation in the marking. It is not known if agreement exists between the markers of the CBM math data and the causes of any disagreement. As the CBM math norming data is currently in use by SD57 it is vital to answer the question about how reliable the markers were. Until inter-rater reliability is determined, the CBM math norming data remains suspect.

This study could determine if gender differences remain an issue of concern. Proven existence of gender differences in reading by Hedekar (1997), and MacMillan (2000) suggested that it may be continuing in SD57. The CBM math research performed by MacMillan (2001) using a Many-Faceted Rasch analysis provided conflicting results regarding the existence of gender differences. However, this research is unconfirmed by conventional analyses. What is undetermined, at present, is whether gender differences were evident in SD57 using the CBM math norming data collected in the 1999-2000 school year. This study can provide evidence, which will confirm or refute the existence of gender differences in math. The need to undertake this study is confirmed by the inconclusive evidence of other researchers regarding gender differences in the field of math. Implementation of this research will answer the question regarding the existence of gender differences in math for SD57. It is unknown if gender differences are specific to the topics of reading and written expression as investigated by Hedekar (1997) and MacMillan (2000). The results of the Many-Faceted Rasch analyses by MacMillan (2001) remain unconfirmed. Therefore, the question of interest is to determine if, within elementary schools in SD57, do gender differences exist in math. If no apparent gender differences in the CBM math exist it is necessary to determine what has changed since Hedekar's study in 1997. It will also be imperative to establish why gender differences

are apparent in CBM reading and writing fluency but not in CBM math. Determining the direction and magnitude of gender differences and how they compare to Hedekar and MacMillan's reanalyses of Hedekar's study will add to our researched understanding regarding their existence. Whatever the outcome, several questions require further investigation.

The research on relative-age has conflicting findings. This study is important to determine if relative-age influences student achievement in math. According to Hedekar (1997) and MacMillan (2000), relative-age was not an issue effecting the academic achievement of students. Boyd (1989) found similar relative-age results. However, Gullo and Burton (1992) state relative-age does play a role in determining academic success. Relative-age research into other subject areas supported the influence of relative-age on academic achievement. Bell and Daniels (1990) conclusions regarding relative-age differences in science provided reason to consider further investigation. MacMillan's (2001) results using Many-Faceted Rasch measurement suggested a lack of relative-age differences in the CBM math data. However, without corroboration of these results, further research is warranted. These inconsistent results indicated there is a need to determine the answer regarding relative-age differences using CBM math. Therefore, it is beneficial for educators, including teachers in SD57, to know if relative-age is worth consideration when academic performance is a concern.

This study contributes to the development of information regarding student performance in SD57. Analyses of the second major CBM norming project in regards to elementary students could highlight new research questions. New questions could include investigation of student achievement in specific sub-groups of the student population.

## CHAPTER THREE: DESIGN AND METHODS

Four separate studies comprise this research. The pre-study examined the inter-rater reliability of the CBM math norming data collected by SD57. Investigation of gender and relative-age differences, effect sizes and effect size comparisons took place in the main study. Finally, two additional studies determined if performance group and grade retention differences existed. This chapter looks at the designs and methods employed to carry out the four studies undertaken in this research.

Within this chapter, investigation of five topics takes place. The first section discusses the subjects of this study. Examination of the instrumentation required for this research is next. The third topic looks at the procedures followed for this study. Then data analysis is explored. Finally, the discussion explores the ethics of this research.

### Subjects

#### *Inter-Rater Reliability*

The term *inter-rater* used for this study is a term used in the literature for a correlation comparison. In this study, the term *marker* refers to the participants of the inter-rater study as they scored the CBM math probes. The markers did not judge or rate the probes.

In this study the subjects for the inter-rater portion were SD57 educators who either attended or presented at the CBM math inservice training held on September 22, 1999. The inservice ensured that before collection of the data from the CBM math probes the educators received training to consistently administer and score the probes. Invitation to volunteer as participants for this study was extended only to educators who attended

the training inservice. Therefore, a convenience sample comprised the selection of participants in the inter-rater study as markers had the choice to take part in the study or not. Three of the 92 educators present at the inservice were presenters and the others were participants. Altogether, 38 markers indicated their agreement to participate by returning the marked inter-rater packages and consent forms to this researcher.

### *Gender and Relative-Age Study*

The main study used the CBM math norming data collected during the 1999-2000 school year. SD57 selected the elementary student subjects for the main study and did not require signed consent forms for student participation. Therefore, this researcher was not required to select the elementary student subjects or collect data for the CBM math research. Out of a population of over 10,000 elementary students, a stratified random sampling of approximately 20% of the population from Grades 1 to 7 was selected. All elementary schools within the district participated in the project, collectively providing 20% of their total student population from each of Grades 1 to 7 as participants.

The CBM Math (Calculation) Norming Project, 1999-2000 hand-outs (see Appendix E) for the training inservice provided information regarding how to randomly select students within each grade and within the school, and the number of students to select from each grade. Therefore, this stratified random sampling process ensured that each grade had equal representation in the norming project. Using a random selection of students provided a range of students in ability, relative-ages and gender. Only students in specific Ministry funded categories were excluded from being chosen for the project. Students were excluded if they were identified as a student classified as Level One or Two English as a Second Language student, a student with a diagnosed Intellectual

Disability, or a student with another “hard label” including: Hearing Impaired, Visually Impaired, Autistic, or Multiply Disabled. The students chosen for this norming project were tested three times during the school year, with the exception of those in Grade 1 who were only tested once in April. To maintain an intact sample of 20% of the population, a procedure was in place to replace students who transferred out during the project.

Walraven and MacMillan (2000) indicate in the Draft Technical Report (see Appendix B) of the CBM (Math) Norming Project that a total of 2039 students were used in the norming sample, representing students in Grades 1 to 7. A break-down of the number of students within each group of the Norming sample is available from the Draft Technical Report (see Appendix B) of the CBM (Math) Norming Project. Out of the total students selected, 48.9% were female and 51.1% were male. Table 1 in the Draft Technical Report of the CBM (Math) Norming Project gives the number of students by grade, also verifying that for the April norming period all grades groups had approximately 14% of the student sample used. This demonstrates that each grade received almost equal representation and consequently, data is available for all grades.

For the purpose of the current study, a further selection took place of students from the district norming sample of 2,039 students. First, only students who participated in all three norming periods of October, January, and April were chosen to participate in the gender, relative-age, performance group and grade retention studies. The exception was Grade 1s who only participated in April. This process eliminated students in Grades 2 to 7 who were missing a score in one or two of the norming periods. The term “elimination of students” refers to the process of eliminating cases from the SPSS CBM

math norming data, which have missing or incorrect data. Incorrect data results from either a data entry mistake, a student who did not meet the criteria for participation or a data case which was removed as part of the data cleaning and screening process. The process of eliminating students ensures the research was not impacted by movement of students into and out of a school or the district. The number of cases removed in each grade for missing data from one or more of the norming periods were 31 Grade 2s, 22 Grade 3s, 45 Grade 4s, 34 Grade 5s, 41 Grade 6s, and 23 Grade 7s. A total of 196 students were eliminated due to missing data for one or two of the norming periods.

Two other students without a birthdate were eliminated. One student was in Grade 5 and the other in Grade 6. At this point, 198 students were eliminated from the data sample.

Next, students who were not the appropriate age for their specific grade were also eliminated from the study, whether they started school early, were retained, or entered school late. A total of 89 cases comprising 5 Grade 1s, 16 Grade 2s, 12 Grade 3s, 20 Grade 4s, 12 Grade 5s, 11 Grade 6s, and 13 Grade 7s were removed due to the inappropriate age for their grade. Of the 89 students removed due to the inappropriate age for their grade, 15 of them had already been eliminated for missing data from one or more of the norming periods.

The procedures to eliminate students either missing data or not the appropriate age for their grade were consistent with those utilized by Hedekar (1997). Thus, comparison of the results is possible between the two studies. Further selection took place with screening and cleaning of the data.

*Data cleaning and screening.* Before data analysis took place, the data was screened to determine that it met the requirements for univariate and multivariate analysis. Descriptive statistics including means, and ranges of the CD scores were calculated for each norming period. No case for an inordinately large or small value was evident.

The next step required determining if any univariate outliers (“cases with an extreme value on one variable” [Tabachnick & Fidell, 2001, p. 67]) were evident in the data. Cases, which produced a  $z$  score of 3.29 or greater on the CD score for each norming period, were considered a potential univariate outlier. This  $z$  score value was chosen as Tabachnick and Fidell recommend “cases with standardized scores in excess of 3.29 ( $p < .001$ , two-tailed test) are potential outliers” (p. 67). Analysis of  $z$  scores produced a total 24 cases to question with eight from October CD scores, seven from January CD scores, and nine from April CD scores. Where possible, cases with high  $z$  scores were verified. One case with an incorrect data entry for the October score was corrected. All actual scores were plausible. A few cases with a  $z$  score indicated scores were decreasing with each norming period rather than increasing as anticipated. However, because all scores were possible a decision was made not to remove any univariate outliers.

Analysis was then carried out to search in the data for the presence of multivariate outliers (“cases with an unusual combination of scores on two or more variables” [Tabachnick & Fidell, 2001, p. 67]). Multivariate outliers were found by calculating the Mahalanobis distance. For this data, the Mahalanobis distance was calculated three times using the October, January and April CD scores as the Dependent Variable with five

Independent Variables. Cases were considered multivariate outliers if their Mahalanobis value was larger than a chi-squared value of 20.515 ( $p < .001$ ,  $df = 5$ ). According to Tabachnick and Fidell, this produces a very conservative estimate that a case is probably an outlier. From this analysis, a total of 13 cases were identified as multivariate outliers. Of these 13 cases, nine were identified in two of the three norming periods as multivariate outliers. Examination of the cases identified proved that they did have erratic and unusual behaviours. The cases exhibited unusual patterns of either scores decreasing across norming periods or larger than anticipated score gains from one norming period to the next. Hence, all 13 cases appear to be true multivariate outliers and do not exhibit the expected pattern of behaviour. A reasonable way to deal with multivariate outliers consisting of less than five percent of the sample is to delete them (Tabachnick & Fidell, p. 90). Because 13 cases represents less than five percent of the sample, the decision was made to delete the multivariate outliers from the sample. After eliminating students due to outlier issues a total of 1754 students remained in the study.

#### *Performance Group Study*

The same 1754 student participants from the main study participated in the performance group differences study.

#### *Grade Retention Study*

Participants in the grade retention study were students who had been eliminated from the main study as they were too old for their grade. However, participants from Grades 2 to 7 took part only if they had data from all three norming periods. A total of 70 students participated comprised of 4 Grade 1s, 11 Grade 2s, 12 Grade 3s, 15 Grade 4s, 11 Grade 5s, 8 Grade 6s, and 9 Grade 7s.



## Instrumentation

### *Inter-Rater Reliability*

Math probes completed by student participants in the SD57 CBM math norming project were used for data for the inter-rater study. From personal experience, the researcher is aware that the Grade 7 probes are more likely than the other grades to provide opportunity for unreliable marking techniques. Therefore, this researcher chose 14 probes from those completed by Grade 7 students for the SD57 math norming research. The probes chosen represented all six probes developed, all three norming periods, a range of student performance, and the most potential for marking discrepancies. After removal of all identifying marks and information from the chosen probes, and darkening of faint answers with a pencil, the probes were photocopied. Each marker received a package with the same 14 probes along with the marking instructions, rules and Grade 7 answer keys. Markers were instructed to mark the probes using the criteria provided at the CBM math norming project inservice.

### *Gender and Relative-Age Study*

Math probes representing the expected learning outcomes appropriate for each grade level were developed for the CBM math norming project of SD57. Teachers who had taught Grades 1 to 7 and possessed a cross-section of knowledge regarding the math curriculum met to develop a bank of math skills for each grade. A school district working group developed six math probes for each grade using skills from the math skills bank (see Appendix D for a sample of a math probe and answer key). A random sampling of skills was chosen for each probe from the grade for which it was developed. Each probe represented the curriculum for the end of the year for each grade. As well, the first three

questions of each probe were drawn from the skills bank of the grade below, and the final three questions were from the grade above.

All probes were administered across each grade and norming period. Therefore, within each testing time gender and relative-age groups for each grade contained all six probes. Due to the distribution of each probe within each gender and relative-age group at each testing time, the differences in difficulty levels between some Grades 5 and 6 probes were not an issue. All other grades were judged equal in difficulty (Walraven & MacMillan, 2000).

The same instrumentation used for the gender and relative-age study was also used for the performance group and grade retention studies.

## Procedures

### *Inter-Rater Study*

A sample of 14 Grade 7 probes formed the marking package. Included in the marking package were the answer keys and marking rules. The following steps were implemented to collect the inter-rater data.

1. The researcher obtained the names of all educators who attended the inservice from Martha Otteson, District Support Teacher, SD57 who co-ordinated the inservice.
2. Email, verbal requests, and personal contact were used to invite educators who attended the inservice to participate in an inter-rater reliability study.
3. Educators agreeing to participate completed a consent form (see Appendix F).
4. The marking packages were sent via school district mail, or personally delivered to each educator who volunteered.

5. Markers were requested to mark the probes according to the CBM math inservice criteria.
6. After marking the probes, the markers returned them via the school district mail, or in person to the researcher along with their completed consent form.
7. Markers who did not return the marking package within the required time were reminded by either phone or email to return the marked probes and consent form.
8. Upon receipt of each marked package, the marker was assigned a number and referred to by that number throughout the study. No information that could identify a marker or their school was referred to during the research.
9. A total of 38 markers returned their packages containing the 14 marked Grade 7 probes. All markers returned individual packages with the exception of two markers who chose to share the same package but differentiated their marking scores by using different colours.
10. Analysis of the marked probes took place after receiving the 14 marked probes from 38 participating markers. Hence, 432 probes were available for analysis.

#### *Gender and Relative Age Study*

To investigate gender and relative-age differences in math, the data collected for the SD57 norming project was used. The data for the SD57 CBM math norming project was collected using the following procedures. Before the start of data collection, a one-day inservice on September 22, 1999 was held to train the elementary teachers and administrators who were to administer the probes in each school. Two educators from each school were invited to attend, although some schools only sent one participant. The three presenters trained 89 educators during the CBM math inservice. When the inservice

was completed a total of 92 educators in the district were trained to administer and score CBM math probes. The hand-outs from the CBM Math (Calculation) Norming Training Project, 1999-2000 provided information for the selection of students, the probes to administer, scoring of the probes and the data collection process (Appendix E).

Following each norming period, the probes were marked and checked by the educators trained during the inservice. Results were then entered on computers at each school into a CBM math template developed for the FileMakerPro version 4.1 program of FileMaker, Inc. After verification of the CBM math probe scores by school personnel, the data was sent electronically to the school board where the data was collated into one file. This file was then sent electronically to Gail Walraven, Master of Education student and Zone Vice-Principal, SD57. To produce the norms for the school district the data was analyzed by Walraven using the SPSS program. Information regarding the norming project was presented at an inservice on September 18, 2000 to teachers in SD57. (For information regarding the Draft Norms Tables for CBM Math Calculation see Appendix C.)

Before analyzing the data, multivariate analysis issues were addressed. First, a decision was made not to investigate normality and linearity as the solution would transform a variable thus making it difficult to interpret the results. The final consideration made was to verify if multicollinearity was apparent within the data. An accurate correlation between variables is required for multivariate analysis. If variables are too highly correlated ( $= .90$  or above) then they are too similar and not all the variables are required for the analysis (Tabachnick & Fidell, 2001, p. 82-83). Examination of the correlations of CD scores between norming periods (found in

Appendix B) indicates the variables have high positive values (.53 to .74). However, none of the values were .90 or higher, so further analysis was possible.

Data analysis on the 1754 cases remaining in the CBM math data took place using the SPSS version 10.1 for Windows statistical package and program.

#### *Performance Group Study*

The procedures for the main study also apply to the performance group study. However, for analysis of performance group differences a new variable was created. For Grades 2 through 7 a variable (CDTotal) was created by totalling the CD scores from the three norming periods. The CDTotal scores for Grade 1 students was their original April CD scores. Percentile ranks for the scores within grades were calculated. Computing the new variable into three percentile groups allowed for the comparison of the high, average and low performance groups. (Refer to the definitions for an explanation of the three performance groups.) These three percentile groups made it possible to compare the means of the performance groups.

As with the main study data analysis took place using the SPSS version 10.1 for Windows statistical package and program on the 1754 cases in the CBM math data.

#### *Grade Retention Study*

Procedures for grade retention followed many of the same procedures as the main study. Retained students were one year older than their appropriate age peers for their same grade level. Three students who were one year ahead of their peers (advanced) were also removed at the same time from the main analyses. As with the gender and relative-age study, the data set for retained students eliminated anyone who was absent for one or more of the norming periods. Retained and advanced students were removed from the

data set before performing univariate and multivariate analyses. The data screening procedures did not apply to the students who were the inappropriate age for their grade. No students were eliminated from the grade retention study due to outlier behaviour as the sample was already small and these students did not meet requirements for the main study. Therefore, analysis took place on the 70 students in this group.

## Data Analysis

### *Inter-Rater Reliability*

Analysis took place once the 38 markers returned their marked probes. Initially, the researcher checked to determine if each marker calculated a CD score for each probe. If the marker had not, the researcher then calculated the CD score for each probe marked, by adding together the score for each question marked on a probe. The first analysis undertaken determined the marking consistency between the markers of the SD57 CBM math norming project by observing the CD scores calculated by the markers for each of the 14 probes. Following that, a Pearson's correlation coefficient determined the inter-rater reliability. Calculation of the mean, standard deviation, coefficient of variation (CV), and upper and lower quartile ranges of the CD scores the markers obtained for each of the 14 probes to determine marker consistency. The coefficient of variation is a dispersion index calculated to allow comparisons of standard deviations to means when the means are markedly different. The coefficient of variation is calculated as  $CV = 100(SD/M)$  (Kirk, 1990, p. 123). Then the same statistics were calculated for each probe. Next, the researcher determined how many markers made addition errors in totalling their CD scores. Finally, comparison of individual question scores was undertaken to discern

the types of questions that produced scoring disagreement among the markers. This final comparison also looked at the range of scores received for specific questions.

### *Gender and Relative-Age Study*

To determine if gender and relative-age differences exist in the CBM math norming data, the CD scores were analyzed with a 2 (gender) x 3 (relative-age) x 3 (norming period) repeated-measures ANOVA by grade. The within-subject variables were the CD scores for each norming period. Gender and relative-age were the variables measured between the subjects. (Refer to the definitions for an explanation of the three relative-age groups.) A 2 (gender) x 3 (relative-age) ANOVA was used to analyze differences between the Grade 1 students as they only participated in the April norming period.

*Considerations for repeated-measures analysis.* Before performing the data analysis, several issues were investigated. A MANOVA analysis is most successful with “highly negative correlated DVs” (Tabachnick & Fidell, 2001, p. 357). As the dependent variables have a moderate positive correlation indicated from observation of the correlations (found in Appendix B), it was decided a repeated-measures ANOVA would be the best approach. Repeated-measures ANOVA require sphericity for the dependent variables (p. 421). A test for homogeneity of covariance, Box’s M, was undertaken to determine if sphericity does exist between the dependent variables. However, according to Tabachnick & Fidell (2001), the Box M test to determine homogeneity of variance is very sensitive (p. 362). Therefore, results may be suppressed in the analyses.

*Effect size comparisons.* Effect sizes were computed for gender and relative-age differences. Cohen's (1992) effect size index was used to determine if the size of the effect was expected to exist in the population. Calculating Cohen's *d* effect sizes also determined if there were hidden trends in the data. Use of Cohen's effect sizes investigated if nonsignificant results found in the main analyses were due to a lack of effect rather than a lack of power. The use of Cohen's *d* also allowed comparisons among multiple groups at the same time, as it is a unitless measure. Data received from Hedekar and MacMillan's (personal communication, January 30, 2002) reanalyses of the CBM reading and writing fluency study was compared to the CBM math data for gender, and relative-age across each grade and norming periods.

Effect size comparisons were undertaken to determine if the gender effects experienced in the CBM math data were consistent with results from other researchers. Where possible the effect sizes were compared to the identical grade. However, if no grade was presented for the data an attempt was made to match the grade. The Grade 7 results were used if the grade was higher or a compilation of results than that measured by the CBM data. As there was data available for all seven grades results from the April norming period were used.

#### *Performance Group Study*

The data was further broken down into the three performance groups for each gender and grade. The performance groups included: a high performance group, an average performance group and a low performance group. This was determined by their achievement on the CBM math norming data (see definitions for further information regarding the performance groups). A 2 (gender) x 3 (relative-age) x 3 (performance



group) x 3 (norming period) repeated-measures ANOVA by grade was undertaken for Grades 2 to 7 to determine if gender differences existed in the high, average, or low performance groups when the means were compared. Grade 1 analysis was performed by a 2 (gender) x 3 (relative-age) ANOVA for grade and performance groups with April CD as the dependent variable. Observation of the number of males and females was also performed, and the ratio of females to males was calculated to determine if more males existed in the high and low groups, thus showing more variability.

#### *Grade Retention Study*

Analysis of differences in math achievement between students who were retained in a grade and students who were the correct age for a grade, was accomplished by comparing the means and standard deviations of the retained students with those who were the correct age for their grade. Observation of the means and standard deviations of each group determined who had the larger mean either the retained students or the students who were the appropriate age for their grade. Next, analysis compared the number of male and female students within each group by grade.

#### Ethics

School District No. 57 did not require signed consent from the students participating in their research as this was considered an in school project appropriately related to the math curriculum. SD57, as part of the CBM math norming project, collected the data for the math probe scores prior to this research. To ensure ethical procedures were followed, permission to use the CBM math norming data was obtained following an ethical review by SD57, Prince George, B.C. (see Appendix A). Permission

for this research project was also obtained from UNBC (also in Appendix A).

Confidentiality was maintained for participants in all aspects of this study. Dr. Peter MacMillan of UNBC will maintain the CBM math norming data, related data and inter-rater analysis files in a secured database. For purposes of further research, information, data files, and inter-rater data will remain the property of Dr. Peter MacMillan. However, the CBM math probes collected by the SD57 for the original CBM math norming project will be retained as their property. Access to all original data, for the completion of every aspect of this study, will be limited to the present researcher, and the researcher's supervisor. Destruction of scored probes will take place upon completion of all research including publications.

## CHAPTER FOUR: RESULTS

Two main topics exist within this chapter. The first topic reports the findings of the inter-rater study. Then the second topic describes the results of the main study. There are three aspects of the main study. Gender and relative-age are the primary focus. Performance group differences are the second section. Finally, grade retention is the last section.

### Inter-Rater Reliability Findings

The CD (correct digit) scores reported for each of the 14 Grade 7 probes are available in Table 1. Differences in severity between markers can be observed from examination of Table 1. Marker 19 is an example of a severe marker by the CD score of zero given to both Probes 8 and 10. A score of zero for both probes is notably discrepant from other markers by a minimum of 29 and 43 marks respectively. The CD score of 99 given by Marker 17 for Probe 5, is higher than any other score by 47 marks. Marker 15 produced another discrepant score with a CD score of 92 for Probe 7, which is 18 marks lower than given by another marker. The CD score of 29 given by Marker 36 is the second lowest score for Probe 8. There is a difference of 19 marks between the score given by Marker 36 and the third lowest score. As indicated, some marking differences were apparent between the markers who participated in the inter-rater study. From the table of CD scores the researcher performed an inter-rater correlation.

Examination of the correlation in Table 2 indicated a high correlation existed between the markers of the CBM math norming project. The mean of the correlation was .98. A Fischer's Z transformation was used to reduce underestimation of the correlation

Table 1

Correct Digit Scores for Inter-Rater Reliability Probes

M	Probe 1	Probe 2	Probe 3	Probe 4	Probe # 5	Probe 6	Probe 7	Probe 8	Probe 9	Probe 10	Probe 11	Probe 12	Probe 13	Probe 14
1	116	28	59	17	45	66	122	67	25	46	57	19	95	63
2	108	31	58	22	46	67	124	61	26	48	56	19	96	66
3	115	31	59	23	47	64	122	66	27	43	58	20	96	66
4	115	33	57	22	49	72	122	63	23	45	59	20	96	66
5	121	31	52	20	48	66	122	67	23	47	50	19	95	56
6	113	31	57	22	49	69	122	67	25	48	59	20	95	64
7	113	31	58	22	47	68	115	53	24	46	57	20	96	66
8	115	31	58	21	46	70	122	65	24	46	57	20	95	61
9	114	31	59	21	49	67	124	67	26	46	56	20	95	75
10	113	34	61	22	51	69	122	67	26	46	59	20	96	64
11	110	28	55	22	45	69	122	48	15	47	59	19	78	66
12	113	24	56	22	48	68	124	66	26	47	57	20	96	65
13	113	31	57	22	46	69	124	68	27	46	58	20	89	63
14	115	34	58	15	48	69	122	67	24	47	58	19	94	60
15	113	31	64	23	52	67	92	60	21	45	57	17	96	64
16	115	31	59	21	49	69	122	67	26	47	53	20	96	65
17	112	31	60	21	99	69	122	67	25	46	59	20	96	58
18	115	31	59	22	49	70	122	65	25	46	54	19	95	62
19	106	22	51	21	37	66	122	0	20	0	59	20	91	62
20	113	30	54	21	44	70	110	65	23	46	57	18	87	64
21	113	33	60	19	40	65	122	57	25	47	59	15	95	63
22	113	32	60	22	51	69	124	56	25	47	59	21	96	64
23	112	32	62	22	42	67	122	66	26	46	58	18	98	63
24	114	33	59	22	48	77	122	66	27	47	58	20	95	65
25	116	33	59	21	45	70	124	63	26	46	57	22	97	69
26	114	33	63	23	46	72	123	63	27	47	58	21	95	65
27	113	30	59	22	50	69	122	67	25	46	58	19	95	62
28	113	33	56	21	46	72	124	54	24	46	59	18	94	62
29	110	23	56	22	47	61	121	66	23	47	55	20	94	66
30	113	30	59	22	45	66	123	65	25	47	56	18	82	64
31	113	31	58	21	47	67	122	67	25	46	57	20	93	64
32	113	29	61	23	44	69	116	68	26	46	55	19	89	64
33	108	31	57	21	49	69	122	48	25	46	57	19	74	63
34	107	32	59	19	42	69	122	57	22	46	59	20	95	60
35	114	31	58	25	45	60	121	61	25	43	58	16	95	57
36	114	31	61	20	46	66	122	29	25	45	54	20	95	64
37	113	31	60	22	49	67	124	67	28	48	57	20	96	66
38	113	31	58	22	45	67	119	68	24	48	56	19	93	62

Table 2

Inter-Rater Correlation Coefficients

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	-												
2	1.00	-											
3	1.00	1.00	-										
4	1.00	1.00	1.00	-									
5	0.99	0.99	0.99	0.99	-								
6	1.00	1.00	1.00	1.00	0.99	-							
7	0.99	0.99	0.99	1.00	0.98	0.99	-						
8	1.00	1.00	1.00	1.00	0.99	1.00	0.99	-					
9	0.99	1.00	1.00	0.99	0.99	1.00	0.99	0.99	-				
10	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	-			
11	0.97	0.98	0.97	0.98	0.97	0.98	0.98	0.98	0.98	0.98	-		
12	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.98	-	
13	1.00	0.99	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.98	1.00	-
14	1.00	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.97	0.99	1.00
15	0.97	0.96	0.97	0.97	0.96	0.97	0.98	0.97	0.96	0.97	0.94	0.96	0.95
16	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.98	1.00	1.00
17	0.91	0.90	0.91	0.91	0.91	0.92	0.90	0.91	0.91	0.92	0.89	0.91	0.90
18	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.98	1.00	1.00
19	0.85	0.87	0.86	0.87	0.83	0.85	0.89	0.86	0.86	0.86	0.89	0.85	0.85
20	0.99	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.99	0.99	0.98	0.99	0.99
21	0.99	1.00	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99
22	0.99	1.00	0.99	1.00	0.99	0.99	1.00	1.00	0.99	1.00	0.99	0.99	0.99
23	1.00	1.00	1.00	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.97	0.99	0.99
24	1.00	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.98	0.99	1.00
25	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
26	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.98	1.00	1.00
27	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.98	1.00	1.00
28	0.99	1.00	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99
29	1.00	0.99	1.00	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.97	1.00	0.99
30	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99	1.00
31	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.98	1.00	1.00
32	1.00	0.99	1.00	0.99	0.99	1.00	0.99	1.00	0.99	1.00	0.98	0.99	1.00
33	0.97	0.98	0.97	0.98	0.97	0.98	0.98	0.98	0.97	0.98	0.99	0.97	0.98
34	0.99	1.00	0.99	1.00	0.98	0.99	0.99	1.00	0.99	0.99	0.98	0.99	0.99
35	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.99	0.99
36	0.95	0.96	0.96	0.96	0.94	0.95	0.98	0.96	0.95	0.96	0.97	0.95	0.95
37	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.98	1.00	1.00
38	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.98	1.00	1.00

Table 2 Continued

	14	15	16	17	18	19	20	21	22	23	24	25	26
1													
2													
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14	-												
15	0.97	-											
16	1.00	0.97	-										
17	0.92	0.90	0.91	-									
18	1.00	0.97	1.00	0.92	-								
19	0.84	0.83	0.85	0.78	0.86	-							
20	0.99	0.97	0.99	0.90	0.99	0.84	-						
21	0.99	0.96	0.99	0.89	0.99	0.87	0.99	-					
22	0.99	0.96	0.99	0.92	1.00	0.89	0.99	1.00	-				
23	0.99	0.97	1.00	0.89	1.00	0.86	0.99	1.00	0.99	-			
24	1.00	0.97	1.00	0.91	1.00	0.86	1.00	0.99	0.99	0.99	-		
25	0.99	0.97	1.00	0.90	1.00	0.87	0.99	1.00	1.00	1.00	1.00	-	
26	1.00	0.97	1.00	0.90	1.00	0.87	0.99	1.00	1.00	1.00	1.00	1.00	-
27	1.00	0.97	1.00	0.92	1.00	0.85	0.99	0.99	0.99	1.00	1.00	1.00	1.00
28	0.99	0.96	0.99	0.90	0.99	0.89	0.99	1.00	1.00	0.99	0.99	1.00	1.00
29	0.99	0.96	1.00	0.91	0.99	0.84	0.99	0.99	0.99	0.99	0.99	0.99	0.99
30	0.99	0.95	0.99	0.90	0.99	0.85	0.99	0.99	0.99	0.99	0.99	0.99	0.99
31	1.00	0.96	1.00	0.91	1.00	0.85	0.99	0.99	0.99	1.00	1.00	1.00	1.00
32	0.99	0.97	1.00	0.90	1.00	0.84	1.00	0.99	0.99	1.00	1.00	0.99	1.00
33	0.97	0.93	0.97	0.90	0.98	0.89	0.97	0.98	0.99	0.97	0.98	0.98	0.98
34	0.99	0.96	0.99	0.89	0.99	0.88	0.99	1.00	1.00	1.00	0.99	1.00	1.00
35	0.99	0.96	0.99	0.91	0.99	0.87	0.99	0.99	0.99	1.00	0.99	0.99	0.99
36	0.95	0.93	0.95	0.87	0.96	0.94	0.94	0.97	0.98	0.95	0.96	0.96	0.97
37	1.00	0.96	1.00	0.91	1.00	0.85	0.99	0.99	0.99	1.00	1.00	1.00	1.00
38	1.00	0.97	1.00	0.90	1.00	0.84	1.00	0.99	0.99	1.00	1.00	1.00	1.00

Table 2 Continued

	27	28	29	30	31	32	33	34	35	36	37	38
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												
26												
27	-											
28	0.99	-										
29	1.00	0.99	-									
30	0.99	0.99	0.99	-								
31	1.00	0.99	1.00	1.00	-							
32	1.00	0.99	0.99	1.00	1.00	-						
33	0.98	0.99	0.97	0.99	0.98	0.97	-					
34	0.99	1.00	0.99	0.99	0.99	0.99	0.98	-				
35	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.99	-			
36	0.95	0.98	0.95	0.95	0.95	0.95	0.97	0.97	0.96	-		
37	1.00	0.99	1.00	0.99	1.00	1.00	0.98	0.99	0.99	0.96	-	
38	1.00	0.99	1.00	0.99	1.00	1.00	0.97	0.99	0.99	0.95	1.00	-

due to the skewness of the sample (Glass & Hopkins, 1996, p. 362). The median of the correlation was .99. The high positive values obtained for the median and mean of the correlation indicate a high rate of agreement between the markers. Most markers produced correlations of .97, .98, .99 and 1.00. The correlation values of 1.00 are actually values of .99 rounded to two decimal places. As indicated by the high positive correlation results, the majority of the markers agreed with the scores given by the other markers indicating a strong relationship between markers. However, the range of correlation scores between all 38 markers varied from a low of .78 to a high of 1.00. Markers 17 and 19 produced the lowest correlation value of .78. Both of these markers produced discrepant scores noted in Table 1. Markers 19 and 17 are responsible for many of the lowest correlation values. For Marker 36 the lowest correlation happens when correlated to Marker 17 with a correlation of .87. Marker 36 also produced discrepant scores as noted in Table 1. All of the low correlations are a result of markers who produced discrepant CD scores compared to others markers. Markers who did not produce CD scores substantially different from the other markers produced the higher correlations observed in Table 2. As the intent of the probes selected was to produce the most variability, this noted range in correlations should reflect the maximum possible variance experienced between markers.

Further patterns of discrepancy in severity and leniency observed between the markers are available in Table 3. From the CD scores the means, standard deviations, coefficient of variation, and quartile ranges of the markers were calculated. Examination of the means showed the means were not identical but varied almost 20 CD from the smallest mean to the largest. Standard deviations varied over nine CD from the smallest



Table 3

Inter-Rater Means and Standard Deviations

Marker	Marker Mean	Marker SD	Marker CV	Marker Q <sub>1</sub>	Marker Q <sub>3</sub>
15	57.29	29.07	50.74	34.50	66.25
10	60.71	31.92	52.57	37.00	68.50
26	60.71	32.07	52.81	36.25	70.25
24	60.93	32.23	52.90	36.50	74.25
32	58.71	31.09	52.96	32.75	68.75
37	60.57	32.24	53.23	35.25	67.00
17	63.21	33.67	53.26	34.75	89.25
20	57.29	30.54	53.32	33.50	68.75
6	60.07	32.12	53.46	35.25	68.50
13	59.50	31.96	53.71	34.75	68.75
7	58.29	31.37	53.83	34.75	67.50
22	59.93	32.27	53.85	35.75	67.75
33	56.36	30.38	53.90	34.75	67.50
38	58.93	31.78	53.93	34.50	67.75
2	59.14	31.93	53.99	34.75	66.75
27	59.79	32.30	54.03	34.00	68.50
9	60.71	32.82	54.06	34.75	73.00
31	59.36	32.12	54.12	34.75	67.00
3	59.79	32.36	54.13	34.00	66.00
16	60.00	32.56	54.26	35.00	68.50
4	60.14	32.64	54.27	36.00	70.50
25	60.57	32.90	54.31	36.00	69.75
30	58.21	31.67	54.40	33.75	65.75
23	59.57	32.53	54.61	34.50	66.75
18	59.57	32.54	54.62	34.75	68.75
8	59.36	32.67	55.04	34.75	68.75
34	57.79	32.02	55.42	34.50	66.75
12	59.43	33.00	55.53	31.25	67.50
14	59.29	32.99	55.64	37.25	68.50
29	57.93	32.29	55.73	29.00	66.00
28	58.71	32.75	55.77	36.25	69.50
35	57.79	32.29	55.87	34.00	60.75
1	58.93	33.39	56.66	32.25	66.75
21	58.07	32.93	56.71	34.75	64.50
11	55.93	31.97	57.17	32.25	68.25
5	58.36	33.77	57.86	35.00	66.75
36	56.57	33.46	59.15	29.50	65.50
19	48.36	38.27	79.14	20.25	65.00
Mean	58.84	32.39	55.18	34.04	68.33
SD	2.26	1.33	4.29	2.89	4.16
CV	3.85	4.11	7.78	8.48	6.08
Q <sub>1</sub>	58.11	31.96	53.74	34.00	66.75
Q <sub>3</sub>	59.98	32.80	55.61	35.19	68.75

to the largest value. Differences in markers are evident from calculation of the standard deviation of standard deviations. If all markers marked identically, the standard deviation would be zero. However, this is not what happened. Because selection of the probes was made with the intent that the means would be different, the coefficient of variation is an appropriate statistic to compare the size of the standard deviations to their means. Within Table 3, the coefficient of variation from the smallest value to the largest ranks the data. This permits expedient observation and comparison of how the marker means and standard deviations differed. If marker severity remained consistent, the researcher expected that the largest mean would produce the largest standard deviation and vice versa for the smallest mean. Observation of the coefficient of variation for each marker showed that Marker 19 with the smallest mean produced the largest coefficient of variation. The second highest coefficient of variation value was from Marker 36, who produced the fourth lowest mean. Marker 17 who had the largest mean did not produce the largest coefficient of variation value. For Marker 17, the coefficient of variation value was smaller than the mean of the coefficient of variation values and the  $Q_1$  of the coefficient of variation values. These results are not what is expected or desired. If the markers were equally severe in their marking then the coefficient of variation value would be identical. The use of quartile ranges eliminated outlier scores from the range values, thus providing a range that reflects the majority of markers. Quartile ranges also assisted in identifying problematic markers. However, eliminating outlier scores still produced a wide range of scores for probes. Marker 17 produced the largest  $Q_3$  score of 89.25 compared to most other markers whose scores varied from 66 to 69. This suggests Marker 17 produced exceptionally large scores compared to the other markers. The

smallest  $Q_1$  score of 20.25 was from Marker 19. This score was nine CD lower than the second lowest  $Q_1$  score suggesting Marker 19 produced some scores much smaller than other markers. Despite the high correlation mean and median produced by the markers, observations of Table 3 demonstrate differences that existed. However, when the descriptive data between the 38 markers is compared most of the markers produced similar scores with only a few marker exceptions. This is evident as the means and standard deviations only vary from each other by one to three CD indicating a high degree of agreement between the markers existed.

Examination of Table 4 illustrates how the probes differed. It also shows that markers responded differently to each probe. As with Table 3, the data for Table 4 is ranked by the coefficient of variation value providing ease of comparison for differences in mean values in relation to the standard deviation. From observation of the means, it is apparent differences existed between the probes. Probes were chosen to represent a range of students, and norming periods. Therefore, the range of probe means found in Table 4 is expected and desired. While a range of means was expected, the anticipation was the coefficient of variation values would remain consistent. If all the markers were identical in marking each probe then coefficient of variation values would be identical. However, Probes 8, 5, and 10 produced coefficient of variation scores substantially bigger than the other probes. This is an indication that the standard deviation is approaching the size of the mean and therefore the standard deviation is larger than anticipated. Thus, for these three probes the markers showed a large variation in their marking. In contrast, Probe 7 with the largest mean produced the fourth smallest coefficient of variation value and provides evidence that a large mean does not lead to an increase in marker differences.

Table 4

Probe Means and Standard Deviations Calculated by the Markers

Probe	Probe Mean	Probe SD	Probe CV	Probe Q <sub>1</sub>	Probe Q <sub>3</sub>
1	113.00	2.68	2.37	113.00	114.00
11	57.08	1.96	3.44	56.25	58.75
6	68.05	2.95	4.33	67.00	69.00
3	58.32	2.61	4.48	57.00	59.75
7	120.89	5.52	4.57	122.00	122.75
14	63.66	3.24	5.09	62.00	65.00
13	93.26	5.12	5.49	93.45	96.00
12	19.32	1.32	6.82	19.00	20.00
4	21.39	1.69	7.95	21.00	22.00
2	30.63	2.64	8.63	31.00	32.00
9	24.58	2.31	9.39	24.00	26.00
10	45.05	7.59	16.84	46.00	47.00
5	47.92	9.03	18.84	45.00	49.00
8	60.63	12.73	20.99	60.25	67.00
Mean	58.84	4.39	8.52	58.35	60.59
SD	32.02	3.31	6.01	32.31	32.31
Q <sub>1</sub>	34.24	2.38	4.50	34.50	35.75
Q <sub>3</sub>	66.95	5.42	9.20	65.75	68.50

Examination of the difference in Q<sub>1</sub> and Q<sub>3</sub> scores provides a further indication that for some probes markers had different ranges of severity. For the majority of probes the interquartile ranges (IQR) are one or two CD different from each other. However, for Probe 8 the IQR is almost seven CD. This is an indication that for this probe the markers exhibited differences in the severity of their marking. The standard deviation of Q<sub>1</sub> and Q<sub>3</sub> is identical which indicates that despite a lower mean for Q<sub>1</sub>, the markers producing lower scores had more variability in marking. As a lower score indicates a more severe marker, it is evident that the severe markers exhibit more variability than the lenient markers. It was expected that the standard deviation of standard deviations would be zero

if all the markers were identical in their severity. However, as this did not happen, it is further evidence of marker disagreement.

This researcher therefore, proceeded to further analyze the probes to determine possible causes or reasons for the differences that existed in probe scores calculated by the markers. It became apparent that more than one reason could explain why the 38 markers did not calculate the exact same CD scores on each of the 14 probes.

### *Reasons for Differences in Probe Scores*

#### *Addition Errors*

Not all markers returned their marked probes with the CD scores totalled and written in the space provided for it. For the 38 markers who participated, the process of determining the total CD score for each probe was not consistent between all markers. The researcher totalled the probes submitted without either a CD score or row totals for Markers 11, and 36. Correct digit scores were calculated by the researcher for probes with row totals and no CD score by adding each of the row totals for Markers 1, 8, 27, 32, 34, and 37. Not all markers had totalled the score of each row before determining the CD score. However, as long as the probe received a CD score the researcher did not see a reason to add row totals to the probes. The researcher checked the accuracy of the probes she totalled. Analysis for the inter-rater study used the CD scores provided by the markers. The process to determine reasons for marker differences verified the accuracy of the CD scores provided by the markers and determined that more than one marker made an addition error while calculating the CD score or row totals. As the researcher did not correct marker addition errors prior to data entry, they contributed to the existence of

marking differences. The researcher calculated and verified the CD score for Markers 11 and 36. Hence, the probes submitted by these two markers did not contribute to the addition errors found by the researcher. There were potentially 504 probes marked by 36 markers, which could have addition errors in their row totals or CD score. A total of 22 marked probes were returned to the researcher with addition errors, as shown in Table 5. The percentage of probes with addition errors was 4.36%. These addition errors are one cause, which contributed to the marking discrepancies.

Table 5

Impact of Addition Errors by Markers

Marker	Probe	Marker CD Total	CD Error	Corrected CD Total
15	7	92	-30	122
21	8	57	-10	67
20	7	110	-5	115
21	12	15	-4	19
26	8	63	-4	67
8	2	31	-2	33
14	14	60	-2	62
19	1	106	-2	108
31	13	93	-2	95
8	3	58	-1	57
17	9	25	-1	26
17	11	59	-1	60
23	12	18	-1	19
35	7	121	-1	122
28	2	33	+1	32
6	10	48	+2	46
25	14	69	+2	67
38	8	68	+5	63
5	1	121	+10	111
9	14	75	+10	65
24	6	77	+10	67
17	5	99	+50	49

Note. + sign indicates addition error added extra marks to the score  
 - sign indicates addition error deleted marks from the score

The percentage of probes with addition errors was only one aspect to consider when determining the impact of addition errors. Table 5 demonstrates how students' marks increased or decreased from one to 50 CD. The addition errors were not limited to one or two specific probes. Probe 4 was the only probe not impacted by addition errors. Addition errors were not limited to probes receiving only high, medium or low scores. Eighteen of the markers contributed to the addition errors found by the researcher. The impact of addition errors was evident in a variety of probes and a diverse range of student scores. In order to assess the influence of the addition errors the researcher corrected the CD totals for all probes with addition errors. Then recalculation of the correlation found the mean with Fischer's *Z* transformation to be .98. The median of the recalculation was .99. When rounded to two decimal places most of the correlation values were .99 and 1.00. Correcting addition errors increased the similarity of correlation values among the markers. Fewer differences were evident. After correcting addition errors, the correlation values indicated a stronger relationship between the markers. Correcting the addition errors also affected the range of the correlations. The minimum correlation increased from .78 to .83. Therefore, the range of correlation values decreased. Now Markers 19 and 5 were responsible for the new low correlation. Marker 17 was no longer responsible for the low correlation when using the corrected addition errors. Not only did correcting addition errors reduce differences between many of the markers but it also reduced the range in correlation values produced by extreme markers. Addition errors were one factor contributing to marker differences but were not the only cause, as the changes in CD scores do not produce an average correlation of 1.00. The researcher continued investigation to find further causes of marker differences.

### *Question Score Discrepancies*

*Unmarked questions.* Questions left unmarked contributed to discrepancies between markers. The term *unmarked question* refers to a question attempted by the student but not scored by the marker with a CD value of zero or larger. A question was considered unmarked if it was without a score, marked with an X, had a line, or a scribble in the question box. The researcher could not accept an “X” as a mark as it was unknown if the marker was following CBM math marking criteria or had reverted to conventional marking practices. CBM math marking requires markers to give credit for partially correct answers. It was unknown if the “X” indicated all the digits were incorrect or if one or more specific digits were incorrect. Therefore, due to the marking process the researcher did not assume that an “X” meant a score of zero. Another reason for not considering unmarked questions a zero score was that other markers did not agree that the score was zero. Therefore, questions were considered unmarked if no CD score was evident.

Analysis of unmarked questions is available in Table 6. When looking through the scoring on each individual probe it was noted that one marker did not mark any questions on the back side of Probe 9. Another marker did not mark any questions on the front side of Probe 8. It is apparent from Table 6 that of the 82 questions given an unmarked score by some markers, only 15 of them received a score of zero by the other markers. The remaining 67 questions, which received an unmarked score by some markers also received at least one score other than zero. Therefore, the majority of unmarked questions (81.70%) added to the discrepancy of scores among the markers. On the 532 probes marked by the 38 markers a total of 7676 questions were to be marked. Out of the 7676



Table 6

Discrepancies in Marking

Probe #	Total # of Questions Attempted by the Student	# of Questions all the Markers Agreed Upon	% of Agreement Between Markers	# of Quest. "Unmarked" by 1 or more Markers But Scored by Other Markers	# of Questions Which Received Either "0" or "Unmarked" by all Markers	Sum of the # of Questions Left "Unmarked" by all Markers
1	25	16	64	4	2	10
2	7	1	14	2	0	2
3	12	5	42	5	0	8
4	4	0	0	2	1	8
5	15	5	33	5	0	19
6	14	10	71	1	0	1
7	19	17	89	0	0	0
8	15	0	0	15	0	40
9	20	5	25	15	7	73
10	11	0	0	11	0	18
11	13	3	23	2	0	16
12	12	1	8	9	3	27
13	18	8	44	4	0	10
14	17	10	59	7	2	57
Total	202	81	40	82	15	289

questions, 289 questions (4%) were not marked. Although this is not a high percent of questions left unmarked it contributes to marking discrepancies.

*Differences in question scores between markers.* Another cause for the differences in CD scores between markers is lack of agreement regarding how many CD a student earned for a question. As shown in Table 6 there were questions for which all the markers agreed on the number of correct digits credited. Obtaining agreement on a question meant that all 38 markers agreed on how many CD a student earned on a question. While this was a stringent requirement, the anticipation was the markers would

not have difficulty reaching agreement. As all markers had received training, they were therefore knowledgeable in the marking process. The CBM math marking rules were enclosed in the marking package so the markers could refer to them if unsure about how to mark a question. In addition, it was expected that most if not all markers had previous experience marking CBM math probes as part of the CBM math norming project or for use in their school. Of the 202 questions attempted on the 14 probes, 81 had marker agreement. Hence, 40% of the questions attempted had agreement between all the markers. Therefore, the remaining 121 questions representing 60% of the questions attempted did not have marker agreement regarding the CD score earned by the student. Table 1 in Appendix G presents analysis of the questions causing marking differences. From this table it is possible to see examples of the types of questions for which markers did not reach scoring agreement. The range of score differences varied from a minimum of one CD to a high of 14. These values provide a look at the impact marker disagreement could have on the CD score a student achieves. Causes of marker disagreement on specific questions are available in Table 7. Some causes of marker differences were due to not following scoring rules. In other cases, causes of marker differences were a result of no specific rule for the markers to follow. When no rules were available, markers differed in the manner they marked the questions. However, despite the existence of several causes contributing to marker differences there were a number of questions for which the markers reached 100% agreement.

Table 7

Causes of Marker Disagreement on Specific Questions

Scoring Rules	Concept or Cause	Cause of Disagreement
	Correct Digits	In a long multiplication question, whether to give credit for each correct digit used to determine the answer, not just the correct answer
# 1	Incomplete Problems	Giving credit for correct digits written for incomplete questions
# 2	Crossed-Out Problems	Giving credit for crossed out problems, in this case the answer was scribbled over but many digits remained visible
# 5	Alignment For Incorrect Answers	If numbers were aligned in a question with a partially correct answer
# 8	Long division	Not giving credit for longest method
# 8	Long division	Giving credit for correct answer when question only partially completed yet answer is correct
#11	Decimals	How to credit answers if decimals are missing or in the wrong location
# 12	Integers	Not giving credit for a (+) sign that is inferred, when the answer is positive
# 12	Integers	Whether to give credit for positive or negative sign when digits incorrect
# 12	Integers	Whether to give credit for digits when positive or negative sign incorrect
No Rules	Marking Concerns	Cause of Disagreement
	Missing digit in ones column	How many digits to credit for an answer missing a zero in the ones column of multi-digit number
	Extra digits in answer	How to mark answers with extra digits, especially when the correct answer is the 2-digits in the hundreds and thousands column followed by two extra zeros, in the ones and tens columns.
	Missing digits in answer	Answers missing a digit, usually a digit in the ones column, the digits written are correct but demonstrate the wrong place value
	Location of Negative sign	Whether to give credit when location of where the negative sign is not directly in front of the digits
	Conversion question	How to credit an answer written as a fraction when converting a decimal to a percent
	Marker Behaviour	Cause of Disagreement
	Unmarked questions	Questions completed or attempted by the students but not marked by marker
	Incorrect credit	Giving credit for incorrect digits

Table 8 presents the concerns that did not cause differences in marking between the 38 markers. For 40% of the questions markers reached agreement on the identical CD score. Hence, some marking concerns did not cause marker disagreement. All markers followed five of the scoring rules. Due to the selection of probes with unusual concerns, it was possible for the researcher to determine which concerns did not cause differences between the markers.

Table 8

Concerns Which Did Not Contribute to Marker Disagreement

Scoring Rules	Concept	Concern
# 3	Regrouping	Rule - No credit for “carries” or “borrows” when regrouping
# 4	Alignment	Alignment for correct answers not required to earn full credit
# 9	Place Holder	“X” used as a place holder counted as a correct digit
# 10	Remainders	Scoring correct remainders only once
# 13	Format	Questions written in the horizontal format where only the answer received credit for correct digits
No Rules	Concept	Concern
	Numbers	Extra numbers or marks written in the computation box, or extra numbers written outside the answer boxes
	Numbers	A zero or X’s added to the front of the correct answer
	Numbers	Messy, poorly formed numerals, or numbers written on an angle
	Answer placement	Answers not written on the answer line provided, when no other issues involved

This researcher found marker reliability to be adequate for the utilization of the CBM math CD scores for the main study. The high correlation mean and median indicated most markers reached agreement on which students earned high marks and which earned low marks. Unanimous agreement on the number of CD earned on 40% of

the questions is a further indicator of marker reliability. Therefore, it was realistic to proceed with analysis of the main CBM math study using the scores calculated by the several markers of the CBM math norming data.

## The Main Studies

### *Gender and Relative-Age Differences*

To determine if a significant difference existed in the analyses alpha was set at 0.01 for both the 2 x 3 x 3 between-subjects repeated-measures ANOVA by grade for Grades 2 to 7 and the 2 x 3 ANOVA Grade 1 students. The analyses were set at a more stringent alpha value (as a Bonferroni-like correction) to compensate for multiple analyses to reduce inflated Type I error (Tabachnick & Fidell, 2001, p. 349). There were six analyses for each of Grades 2 to 7 and a seventh analysis for Grade 1. The homogeneity of the covariance matrix was examined using the Box M test to test for sphericity. The Box M test is known to be very sensitive. With the exception of Grade 2, all other Box M results were nonsignificant. The Grade 2 results produced a significant result  $p < .001$ , therefore for this grade, robustness to violations of assumptions is not guaranteed (p. 330). Hence, Pillai's Trace is the recommended multivariate statistic to use. The Pillai's Trace values were recalculated as converted  $F$  values. Results of the repeated-measures ANOVA of gender and relative-age differences for Grades 2 to 7 are presented in Table 9. Grade 1 results in Table 9 are from the 2 x 3 ANOVA. Besides the multivariate statistic, Table 9 provides the degrees of freedom for error and hypothesis, and significant values of Pillai's Trace.

Table 9

Repeated-Measures ANOVA for Gender and Relative-Age

Grade	Source	F	df	p	
1	Gender	3.56	1	.19	
	RA	1.44	2	.41	
	G*RA	2.25	2	.11	
		F (from V)	df <sub>h</sub>	df <sub>e</sub>	p
2	Gender	0.39	2	242	.70
	RA	1.17	4	486	.33
	G*RA	0.06	4	486	.99
3	Gender	0.43	2	248	.65
	RA	3.99	4	498	.00*
	G*RA	0.85	4	498	.50
4	Gender	0.77	2	230	.47
	RA	0.38	4	462	.82
	G*RA	1.33	4	462	.26
5	Gender	1.02	2	242	.36
	RA	0.88	4	486	.48
	G*RA	1.69	4	486	.15
6	Gender	0.64	2	237	.53
	RA	0.71	4	476	.59
	G*RA	0.52	4	476	.72
7	Gender	0.53	2	236	.59
	RA	1.10	4	474	.36
	G*RA	3.63	4	474	.01*

 $\alpha = .01$  for V and F

Gender differences were not apparent in the CBM math CD score analysis. In fact, gender effects were not evident in the data for any of the seven grades. None of the seven grades came close to demonstrating a significant gender difference. The smallest  $p$  value was .19 followed by .36. Nor, did there appear to be a trend regarding gender differences within the data. Hence, gender differences do not exist in the CBM math data at either primary or intermediate grades according to the repeated-measures ANOVA.

Relative-age differences throughout the data are not consistent. From Table 9 it is evident there is only one grade, which produced significant relative-age differences Grade 3,  $F(4, 498) = 3.49, p = .00$ . The level of significance was actually  $p = .003$ . Even if a more lenient alpha value of .05 were set there would still be only one grade demonstrating a significant relative-age difference. In addition, investigation of the data to discover the existence of a trend in relative-age difference did not produce a trend. One significant result compared to six nonsignificant results indicates that consistent relative-age differences did not exist. Therefore, relative-age differences are not considered evident in the CBM math data for elementary students.

Examination of the interaction between gender and relative-age was also possible from Table 9. Interaction effects were not apparent in Grades 1 to 6. However, Grade 7  $F(7, 474) = 3.63, p = .01$ , produced the only significant interaction. The actual significance level for Grade 7 is  $p = .006$ . This was the only evident interaction in the CBM math data; a consistent pattern of significant interactions was not demonstrated. However, if a cautious approach were taken in the investigation of an interaction a  $p$  value of .25 would be considered. Dividing this cautious  $p$  value by the six analyses undertaken produces a cautious  $p$  value of .04. Even when using a  $p$  value of .04 there remained only the one

significant interaction. The  $p$  value of .11 for Grade 1 remained nonsignificant even though it was a value of interest. To provide further confirmation whether gender and relative-age interactions existed the interactions for Grade 1 and 2 were graphed. Grade 1 was chosen as it has a alpha value of interest. It was decided to determine if the sensitive Box M also signified a significant Grade 2 interaction was hidden in the analysis.

An ordinal interaction was evident for Grade 1 students. The interaction line for females remained almost horizontal for all three relative-ages. However, this pattern did not hold for males. The mean for young males was nearly identical to that of young females. The difference in mean distance between males and females increased as the relative-age changed from young to average to old. Thus, for Grade 1 students there appeared to be a significant interaction.

Inspection of the interaction graphs for Grade 2 students at all three norming times was undertaken. An ordinal interaction was evident in the October Grade 2 graph. For this graph the mean for males and females was furthest apart when students were young and came together slightly for average-aged Grade 2 students. When students were in the oldest age group the means were very similar. In January Grade 2 the graph followed a similar pattern to October. However, the difference between the young students was not as pronounced. By April, Grade 2 the interaction was much less pronounced than for the other two norming periods. The most obvious mean difference between male and female students in April of Grade 2 is for young students. A slight decrease in difference between average and older students was evident. Hence, as with Grade 1 students the graphs for October, January, and April Grade 2 confirm an ordinal



interaction was evident between Grade 2 students. However, this may have been a hidden interaction.

Observation of the data for Grades 3 to 7 indicates a trend was not apparent in the gender and relative-age interactions. The result therefore, is that an interaction between gender and relative-age is apparent for Grade 1 and the first two norming periods of Grade 2. This indicates a hidden interaction was evident for early primary grades but no interaction was evident for the other grades.

#### *Means, Mean Square Within and Effect Sizes*

Means, mean square within, and effect sizes for gender at each grade level and norming period are presented in Table 10. As is evident from the means and the value of Cohen's  $d$  one gender does not consistently produce a higher mean than the other does throughout all grades.

The highest mean in the data was for girls in Grade 7 in April with  $M = 60.52$ . In April, Grade 6 students demonstrated the most variation in scores with a mean square within of 543.77. The highest mean was not accompanied by the most variation in the data. During the October norming periods students in Grade 4 produced the least variation in the gender data with a mean square within of 100.48. The lowest mean in the gender CBM data was produced by Grade 2 girls in October with  $M = 12.16$ . As with the highest mean and mean square within, the lowest mean and mean square within are not from the same gender, grade or norming period. There were however, two patterns evident in the gender effect size data (Table 10). In the primary grades (1 to 3) as well as Grade 5, all the effect sizes favoured boys. The girls were favoured by the effect sizes in the other three grades, which are all the intermediate grades with the exception of Grade

Table 10

Mean CD Scores and Effect Sizes for Grade and Gender by Norming Period

Grade	Norming Period	Gender	Mean	MS <sub>w</sub>	Cohen d
1	April	F	13.24	101.50	-0.36
		M	16.86		
2	October	F	12.16	102.66	-0.37
		M	15.95		
	January	F	23.14	188.26	-0.34
		M	27.75		
	April	F	30.32	228.39	-0.21
		M	33.52		
3	October	F	21.16	148.50	-0.07
		M	22.05		
	January	F	30.72	180.51	-0.08
		M	31.78		
	April	F	37.36	192.95	-0.06
		M	38.16		
4	October	F	20.81	100.48	0.20
		M	18.81		
	January	F	33.61	239.58	0.12
		M	31.76		
	April	F	37.71	302.42	0.06
		M	36.67		
5	October	F	21.34	129.86	-0.15
		M	23.08		
	January	F	34.58	265.06	-0.01
		M	34.73		
	April	F	37.99	345.74	-0.14
		M	40.52		
6	October	F	44.18	352.05	0.03
		M	43.69		
	January	F	56.64	445.24	0.05
		M	55.62		
	April	F	60.57	543.77	0.12
		M	57.81		
7	October	F	46.68	417.49	0.01
		M	46.54		
	January	F	56.50	495.51	0.20
		M	52.15		
	April	F	60.52	509.27	0.22
		M	55.65		

Note: A negative effect size indicates boys are favoured over girls.

5. Another pattern that was consistent throughout all grades is that within each grade, one gender is favoured over another for all three norming periods. As Cohen (1992) describes effect sizes, a trivial effect is below 0.2, a small effect is 0.20 to  $< 0.50$ , a medium effect is 0.5 to  $< 0.80$ , and a large effect is 0.8 or greater. By Cohen's definitions, there are no large or medium effects evident in the gender effect sizes. Small effects are evident in Grade 1, all Grade 2 norming periods, October Grade 4, and January Grade 7. All other effects are considered nonsignificant. A trend is not evident in the effect size data for gender. Evidence of small or trivial effect sizes produces a trivial practical result. These small and insignificant effect sizes confirm the CBM math data does not appear to indicate a lack of gender differences.

The consistent pattern of small gender effect sizes for Grade 1 and all Grade 2 norming tests prompted the researcher to perform further analysis. If a difference exists between genders in Grades 1, 2 as well as January and April of Grade 7 as suggested by the small effect sizes then it is possible the significant results were suppressed in the original analysis. Hence, a  $t$  test for two-independent sample means was performed on a highly similar sample for Grades 1 to 7. This was to verify an actual difference between genders at Grades 1, and 2. It also provided the opportunity to determine if the small effects for the January and April Grade 7 norming periods were significant. With alpha set at  $p < .05$  for a two tailed test, a significant result was evident at Grade 1 with  $p = .003$ , for equal variances assumed. Setting the same conditions for Grade 2 also produced significant results for the October and January norming periods with  $p = .004$  and  $p = .009$ , respectively. A significant  $t$  test result was not evident for any other grades or the April Grade 2 norming period with  $p = .074$ . As the significant  $t$  test results produced for

Grade 1 and the first two norming periods for Grade 2 match the three gender effect sizes, it is evidence that a small but significant gender effect was evident for Grade 1 and the first half of Grade 2. The lack of a significant  $t$  test result for April of Grade 2 as well as January and April of Grade 7 reflects the effect sizes which were only just large enough to be considered a small effect rather than a trivial effect.

Table 11 presents the relative-age means, mean square within, and effect sizes for each norming period. The oldest Grade 6 students in the April norming period produced the largest mean with  $M = 61.34$ . In this instance, the Grade 6 students during April also produced the most variation in the data with a standard deviation of 23.40. Young, Grade 1 students in April had the lowest mean with  $M = 12.56$ . As with the gender data in Table 10 the Grade 4 students in October with a mean square within of 100.48 had the least variation in data.

The effect sizes for relative-age were calculated by subtracting the average group from either the oldest or youngest, then dividing the square root of the mean square within. This is to demonstrate how different the older and younger students were from their average aged peers. As the average group represented the largest group of students, this researcher concluded this was the most typical group and was therefore chosen as the group to compare the oldest and youngest students to. There were no large or medium effect sizes apparent in the relative-age CBM math data. The Grade 7, October effect for young students was the largest effect with  $d = -0.49$ . All other effects are within the magnitude of a small effect or nonexistent. With a few exceptions, the majority of relative-ages had no effect. Those with small effects only minimally met the criteria for an effect of a small magnitude. A trend in relative-age effect sizes was not apparent.

Table 11

Mean CD Scores and Effect Sizes for Grade and Relative-Age by Norming Period

Grade	Norming Period	Relative-Age	Mean	MSw	Cohen d
1	April	Young	12.56	101.50	-0.24
		Average	14.97		
		Old	17.49		0.25
2	October	Young	12.59	102.66	-0.08
		Average	13.43		
		Old	16.89		0.34
	January	Young	23.06	188.26	-0.17
		Average	25.44		
		Old	27.71		0.17
	April	Young	29.31	228.39	-0.22
		Average	32.66		
		Old	32.62		0.00
3	October	Young	18.69	148.50	-0.31
		Average	22.44		
		Old	22.22		-0.02
	January	Young	29.11	180.51	-0.11
		Average	30.58		
		Old	34.35		0.28
	April	Young	35.24	192.95	-0.08
		Average	36.39		
		Old	42.66		0.45
4	October	Young	18.45	100.48	-0.15
		Average	19.93		
		Old	20.35		0.04
	January	Young	31.74	239.58	-0.04
		Average	32.43		
		Old	33.60		0.08
	April	Young	37.64	302.42	0.06
		Average	36.65		
		Old	37.80		0.07

5	October	Young	21.71	129.86	0.01
		Average	21.58		
		Old	24.12		
	January	Young	32.60	265.06	-0.16
		Average	35.20		
		Old	36.14		
	April	Young	36.42	345.74	-0.19
		Average	40.03		
		Old	41.24		
6	October	Young	41.77	352.05	-0.12
		Average	44.08		
		Old	45.80		
	January	Young	52.66	445.24	-0.22
		Average	57.21		
		Old	57.58		
	April	Young	53.82	543.77	-0.31
		Average	61.02		
		Old	61.34		
7	October	Young	38.13	417.49	-0.49
		Average	48.20		
		Old	51.46		
	January	Young	48.80	495.51	-0.25
		Average	54.26		
		Old	59.98		
	April	Young	53.15	509.27	-0.24
		Average	58.62		
		Old	61.21		

Therefore, relative-age effects are minimal and were not an issue. As with gender effect sizes the statistically nonsignificant results were due to trivial effects and a trivial practical significance. Trivial effect sizes for relative-age were reflected in the CBM math data repeated-measures ANOVA analysis.

As with the gender data, the researcher decided to perform *t* tests on the relative-age data. This was to determine if effects were suppressed in the original analyses. To verify if there is an actual difference between old and average aged students at all seven

grades a  $t$  test for two-independent sample means was performed on a highly similar sample. The same analysis was completed for young and average aged students. Alpha was set at  $p < .05$  for both analyses. Significant results were evident for old and average aged students in October Grade 2 with  $p = .032$  and April Grade 3 with  $p = .004$  for equal variances assumed. For young and average age students significant results were evident for April Grade 6 with  $p = .046$  and October Grade 7 with  $p = .002$ . The  $t$  tests only produced significant relative-age results for substantial small effect sizes. Significant  $t$  test results were not produced for small effect sizes that were not well within the small category. The lack of pattern and consistency in the significant  $t$  test results further confirms prior evidence suggesting relative-age differences do not exist.

### *Effect Size Comparisons*

*Comparison with CBM reading and writing.* In order to determine if gender and relative-age effects for CBM math were identical to those for CBM reading and writing a comparison is presented in Tables 12 and 13 respectively. Hedekar and MacMillan (2002) (personal communication, January 30, 2002) calculated the data for the CBM reading and writing effect sizes. Using Cohen's (1992) effect size values, effect sizes were considered to have changed magnitude if they moved from one size category to another. Recall that effect sizes below 0.20 were considered trivial, those from 0.20 to < 0.50 were small, from 0.50 to < 0.80 was a medium effect and values larger than 0.80 were large.

Examination of the comparison of effect sizes for CBM math and reading indicates the magnitude of the effect changed for ten of the nineteen comparisons. The changes in magnitude were not consistent for the Grades 3, 4 and 7 across all three

Table 12

CBM Math and Reading Gender Effect Size Comparison

Grade	Norming Period	CBM Math (CD)	CBM Reading (WRC)	Magnitude Change	Direction Change
		Cohen d	Cohen d		
1	April	-0.36	0.27	No	Yes
2	October	-0.37	0.49	No	Yes
	January	-0.34	0.42	No	Yes
	April	-0.21	0.47	No	Yes
3	October	-0.07	0.17	No	Yes
	January	-0.08	0.26	Yes	Yes
	April	-0.06	0.28	Yes	Yes
4	October	0.20	0.37	No	No
	January	0.12	0.32	Yes	No
	April	0.06	0.24	Yes	No
5	October	-0.15	0.56	Yes	Yes
	January	-0.01	0.41	Yes	Yes
	April	-0.14	0.45	Yes	Yes
6	October	0.03	0.25	Yes	No
	January	0.05	0.33	Yes	No
	April	0.12	0.31	Yes	No
7	October	0.01	0.25	Yes	No
	January	0.20	0.19	No	No
	April	0.22	0.30	No	No

Note: A negative effect size indicates boys are favoured over girls.

CBM Reading was calculated as Words Read Correctly (WRC).

CBM Reading (WRC) effect size received from L. Hedekar & P. MacMillan (personal communication, January 30, 2002).



Table 13

CBM Math and Writing Gender Effect Size Comparison

Grade	Norming Period	CBM Math (CD)	CBM Writing (WSC)	Magnitude Change	Direction Change
		Cohen d	Cohen d		
1	April	-0.36	0.47	No	Yes
2	October	-0.37	0.59	Yes	Yes
	January	-0.34	0.52	Yes	Yes
	April	-0.21	0.69	Yes	Yes
3	October	-0.07	0.36	Yes	Yes
	January	-0.08	0.48	Yes	Yes
	April	-0.06	0.56	Yes	Yes
4	October	0.20	0.49	No	No
	January	0.12	0.47	Yes	No
	April	0.06	0.66	Yes	No
5	October	-0.15	0.70	Yes	Yes
	January	-0.01	0.56	Yes	Yes
	April	-0.14	0.68	Yes	Yes
6	October	0.03	0.62	Yes	No
	January	0.05	0.63	Yes	No
	April	0.12	0.72	Yes	No
7	October	0.01	0.55	Yes	No
	January	0.20	0.48	Yes	No
	April	0.22	0.60	Yes	No

Note: A negative effect size indicates boys are favoured over girls.

CBM writing was calculated as (WSC).

CBM writing effect size received from L. Hedekar & P. MacMillan (personal communication, January 30, 2002).

norming periods. No change in magnitude was experienced by Grades 1 and 2. Grades 5 and 6, tested at the three norming periods, remained consistent in the existence of a change in magnitude. Changing effect size magnitude was not consistent when comparing CBM math and reading, between norming periods or from grade to grade.

However, the change in direction of effect size was consistent throughout all norming periods within a grade. For Grades 1 to 3 and Grade 5 the effect size consistently changed from favouring girls in reading to favouring boys in CBM math. Grades 4, 6, and 7 demonstrate a consistent pattern for CBM reading and math in that the effect favoured girls. In comparison to the CBM reading effects, which consistently favoured girls and with most effects were a medium size in magnitude, the CBM math effects were not consistent in direction nor did they demonstrate the same magnitude of effect size. CBM reading and math results were therefore not producing the same gender effects.

The results from comparing the CBM math and writing (WSC) were more dramatic than the CBM reading and math. With the exception of the Grade 1 students, and the Grade 4 October results, all other comparisons showed a dramatic change in magnitude of effect. This was demonstrated by the Grade 4 April results which had an extremely small or nonexistent effect size of 0.06 in CBM math, yet had a medium effect in CBM writing of 0.66. When comparing the direction of the effect between CBM math and writing, there was no change for Grades 4, 6, and 7. For the other grades the direction changed from CBM writing to math. In CBM math, the effect did not consistently favour one gender over another. However, in CBM writing girls were always favoured and with a substantial difference in effect. Comparison of CBM writing and math effects indicated gender differences were very different between the two subjects. While there were few effects evident in the CBM math data, the opposite was apparent in the CBM writing results.

*Comparison with other math research.* Table 14 presents the results of the CBM math effect size comparison to other research. Comparison of gender effect sizes from

Table 14

Gender Effect Size Comparison for CBM Math and Other Math Research

Grade	CBM Math Cohen d	Other Math Research			Researcher	Age or Grade
		Cohen d or Standard Mean Difference	Magnitude Change	Direction Change		
1	-0.36					
2	-0.21					
3	-0.06					
4	0.06	-0.04	No	Yes	Beller & Gafni (1996)	9 years
4	0.06	<0.10	No	No	Cole (1997)	Gr. 4
5	-0.14					
6	0.12					
7	0.22	-0.12	Yes	Yes	Beller & Gafni (1996)	13 years
7	0.22	-0.11	Yes	Yes	Beller & Gafni (2000) (1991 data)	13 years
7	0.22	-0.03	Yes	Yes	Beller & Gafni (2000) (1988 data)	13 years
7	0.22	-0.03	Yes	Yes	Hedges & Nowell (1995) p. 43	Gr. 8
7	0.22	-0.11	Yes	Yes	Willingham & Cole (1997) p. 122	Gr. 12
7	0.22	<-0.10	Yes	No	Cole (1997)	Gr. 12
7	0.22	0.18	Yes	No	Kleinfeld (1998a) Math Computation	Gr. 12
7	0.22	-0.11	Yes	No	Kleinfeld (1998a) Math Concepts	Gr. 12

Note: A negative effect size indicates boys are favoured over girls.

other research with the CBM math indicates effect sizes for some CBM math results were larger than those determined by other researchers. Beller and Gafni (1996) found the

effects for nine year olds favoured boys whereas the CBM math data favoured girls. However, the size of the effects when compared is almost identical and nonexistent. Whereas the CBM math for Grade 7 favoured girls the effect sizes for 13 year olds from Beller and Gafni favoured boys. While the Grade 7 CBM math effect was considered small in magnitude, the 13 year old effect is trivial. The Grade 12 effect by Willingham and Cole (1997) and the Grade 7 CBM math effects both favour girls. A larger effect magnitude is found in the CBM data, in the small range compared to the Grade 12 data, which has a trivial effect size magnitude. If Grade 6 or October Grade 7 CBM effect sizes were used for comparison instead of April Grade 7 effect sizes, no changes in magnitude would exist between the CBM data and other researchers. Even though CBM math effect sizes are small or trivial they appear to be slightly larger than the effect sizes found by several other researchers. Direction of effect sizes was not consistent in two cases with the CBM math data. However, the CBM data does not demonstrate a consistent effect size direction.

### *Performance Group Differences*

The repeated-measures ANOVA analysis for the low and high performance groups are presented in Tables 15 and 16 respectively. Results for the average performance group are available in Appendix H. Along with performance group values, relative-age values are also presented. The issues regarding the repeated-measures ANOVA for gender and relative-age also applied to this analysis. As with the gender and relative-age data, the homogeneity of the covariance matrix was examined using the Box M test for sphericity. The Grade 6 average performance group produced a significant Box

Table 15

Repeated-Measures ANOVA for the Low Performance Group

Grade	Source	F	df		p
1	Gender	0.04	1		.86
	RA	0.86	2		.54
	G*RA	1.08	2		.35
		F (from V)	df <sub>h</sub>	df <sub>e</sub>	p
2	Gender	0.92	2	56	.41
	RA	0.41	4	114	.80
	G*RA	0.20	4	114	.94
3	Gender	4.32	2	57	.02
	RA	1.63	4	116	.17
	G*RA	2.32	4	116	.06
4	Gender	1.51	2	50	.23
	RA	1.23	4	102	.30
	G*RA	1.04	4	102	.39
5	Gender	0.52	2	53	.60
	RA	0.74	4	108	.57
	G*RA	0.55	4	108	.70
6	Gender	0.50	2	52	.61
	RA	0.14	4	106	.97
	G*RA	1.11	4	106	.36
7	Gender	0.41	2	54	.66
	RA	1.09	4	110	.37
	G*RA	0.53	4	110	.71

 $\alpha = .01$  for V and F

Table 16

Repeated-Measures ANOVA for the High Performance Group

Grade	Source	F	df		p
1	Gender	3.65	1		.17
	RA	1.25	2		.45
	G*RA	1.07	2		.35
		F (from V)	df <sub>h</sub>	df <sub>e</sub>	P
2	Gender	0.60	2	57	.55
	RA	2.48	4	116	.05
	G*RA	1.82	4	116	.13
3	Gender	0.14	2	58	.87
	RA	2.63	4	118	.04
	G*RA	0.62	4	118	.65
4	Gender	1.55	2	53	.22
	RA	2.07	4	108	.09
	G*RA	1.04	4	108	.39
5	Gender	0.13	2	54	.88
	RA	1.40	4	110	.24
	G*RA	2.06	4	110	.09
6	Gender	1.23	2	54	.30
	RA	1.26	4	110	.29
	G*RA	0.97	4	110	.43
7	Gender	0.10	2	53	.90
	RA	1.59	4	108	.18
	G*RA	1.26	4	108	.29

$\alpha = .01$  for V and F

M result of  $p = .001$ . With this one exception, all other performance groups produced a nonsignificant ( $p < .001$ ) Box M result. Again, Pillai's Trace is the recommended multivariate statistic to use, since robustness is not guaranteed for one analysis, and to remain consistent between analyses. The alpha was set at a stringent value of  $p < .01$  to

compensate for multiple analyses. The Grade 1 results in Table 15 and 16 are from an ANOVA since Grade 1 only participated during one norming period.

Dividing the CBM math data into performance groups within each grade did not produce significant gender differences. Examination of the alpha values indicated that neither the low nor high performance groups showed existence of gender differences. Nor did the data reveal that a trend in gender differences appeared in the performance group results. Hence, gender differences within performance groups are not an issue.

Further investigation took place to determine if the lack of gender differences within performance groups was impacted by not restricting the high and low performance groups to students in the top or bottom 10th percentile. The additional analysis defined the high performance group above the 90th percentile and the low performance group below the 10th percentile. Once again, for Grades 2 to 7 a 2 (gender) x 3 (relative-age) x 3 (norming period) x 3 (performance group) repeated measures ANOVA with three repeated measures of: October, January, April CD scores by grade and performance groups was carried out. Analysis for Grade 1 students was a 2 (gender) x 3 (relative-age) ANOVA. However, robustness of this analysis was not guaranteed due to the cell sizes below the required 20 (Mardia, cited in Tabachnick & Fidell, 2001) for high and low performance groups. However, no evidence of gender differences was apparent for Grades 1 to 7 when the performance groups were redefined. Results of this additional analysis provided further evidence of a lack of gender differences even when changing the definition of the high and low performance groups.

The investigation into relative-age differences in performance groups was not one of the hypotheses developed for this study. However, the analysis allowed for

investigation of this topic. Therefore, the researcher included the results with this study. However, some cells for the young and older-aged students contained fewer than 20 students. Thus, robustness of this multivariate relative-age analysis within performance groups was not guaranteed (Mardia, cited in Tabachnick & Fidell, 2001). The relative-age differences investigated within the data also reflected no apparent differences. Within the low and high performance groups, no significant relative-age differences were produced. In addition, the data did not demonstrate evidence that a trend in relative-age differences existed. Thus, relative-age differences within performance groups are not apparent.

No evidence of an interaction or a trend was apparent within the data. There were no significant interactions produced from the CBM math data. Therefore, dividing the CBM math data into performance groups did not reveal evidence of main effects for gender or relative-age differences, nor did it find an indication of an interaction between gender and relative-age.

As would be expected, based on the gender and relative-age results, the average performance group did not produce any results contradicting the original gender and relative-age results. No significant gender, relative-age, or interaction results were evident. Therefore, the results are not reported here. Table 1 in Appendix H provides the results of the average performance group.

Comparison of the ratio of female and male students in each performance group is provided in Table 17. The ratios of females to males were determined by dividing the number of females by the number of males within a performance group. From examination of Table 17 the ratios of females to males did not follow any specific pattern



Table 17

Number of Students by Gender and Performance Group

Grade	Gender	High Performance Group		Average Performance Group		Low Performance Group	
		N	Ratio	N	Ratio	N	Ratio
1	F	29	0.67	63	0.83	38	1.41
	M	43		76		27	
2	F	22	0.52	63	1.07	34	1.17
	M	42		59		29	
3	F	30	0.86	72	1.33	30	0.88
	M	35		54		34	
4	F	34	1.31	58	0.94	26	0.84
	M	26		62		31	
5	F	28	0.85	66	1.06	31	1.07
	M	33		62		29	
6	F	36	1.44	69	1.25	27	0.84
	M	25		55		32	
7	F	30	1.00	59	0.94	24	0.65
	M	30		63		37	

Note: A ratio smaller than 1.0 favours males and a ratio larger than 1.0 favours females.

in any of the three performance groups. Nor did a trend in the ratio of females to males develop in the data. Within the performance groups one gender did not consistently outnumber the other. This finding is consistent with the lack of significant gender differences found within the performance group data.

*Grade Retention*

The mean, standard deviation and numbers of retained students and appropriate age students for the same grade are presented in Table 18. Data for the advanced students was not presented, as there were too few students to be a valuable comparison. This researcher does not know the reasons for retention or advanced placement. The sample size for retained students was less than four percent of the size of the sample for

Table 18

Mean CD for Retained and Appropriate Age Students by Grade and Gender

Grade	Norming Period	<u>Retained Students</u>					<u>Appropriate Age Students</u>				
		Mean	SD <sub>totalret</sub>	n <sub>total</sub>	n <sub>f</sub>	n <sub>m</sub>	Mean	SD <sub>totalapp</sub>	n <sub>total</sub>	n <sub>f</sub>	n <sub>m</sub>
1	April	12.75	4.03	4	0	4	15.14	10.36	277	131	146
2	October	12.27	10.18	11	4	7	14.14	10.35	249	119	130
	January	23.27	12.90	11	4	7	25.55	13.88	249	119	130
	April	30.73	14.53	11	4	7	31.99	15.11	249	119	130
3	October	22.67	10.97	12	5	7	21.59	12.18	255	132	123
	January	27.67	12.54	12	5	7	31.23	13.46	255	132	123
	April	32.33	14.25	12	5	7	37.75	14.15	255	132	123
4	October	14.53	11.32	15	6	9	19.80	9.99	237	118	119
	January	22.27	12.94	15	6	9	32.68	15.46	237	118	119
	April	25.40	11.03	15	6	9	37.19	17.29	237	118	119
5	October	22.82	12.57	11	4	7	22.21	11.37	249	125	124
	January	35.27	22.88	11	4	7	34.66	16.35	249	125	124
	April	40.64	26.70	11	4	7	39.25	18.61	249	125	124
6	October	31.38	20.74	8	3	5	43.95	18.80	244	132	112
	January	48.38	20.26	8	3	5	56.17	21.04	244	132	112
	April	44.38	20.88	8	3	5	59.30	23.40	244	132	112
7	October	46.00	19.70	9	1	8	46.60	20.87	243	113	130
	January	48.33	26.23	9	1	8	54.18	22.47	243	113	130
	April	55.11	23.04	9	1	8	57.92	22.84	243	113	130

Note. The symbols f = female, and m = male.

appropriate age students. As the data set for retained students was not screened for univariate or multivariate outliers a possibility exists that the results for retained students were influenced by the presence of one or more outliers. Thus making the mean for the one or more retained grade groups much larger or smaller than the population it represents. The influence of potential outliers in the data for retained students could be compounded by the small sample size within each grade. In fact, none of the cells for

retained students were large enough to produce a robust multivariate analysis (Mardia, cited in Tabachnick & Fidell, 2001). Therefore, further statistical analysis is not a worthwhile undertaking without first increasing the size of the sample.

Comparing means for retained students and their appropriate aged peers in the same grade level did not produce a consistent result for all seven grades. In five grades, the retained students had a lower mean than their appropriate age peers in the same grade level did for each norming period. However, for Grade 5, the retained students produced higher means for CD scores than did their peers. In Grade 3, the October mean was higher for retained students while the mean for January and October was higher for appropriate aged students. Therefore, for Grades 1, 2, 4, 6, and 7 retained students did not score as well as students who were the correct age for their grade. While not a consistent pattern, the results indicate that retained students do not perform as well as students placed in the correct grade for their age.

A pattern did develop in the retained data when comparing the number of male and female students. For each grade, there were more males than females in the retained group. In fact, the number of males in each grade was more than twice the number of females. At Grade 1, only males composed the retained group. The ratio of males to females in the retained group is quite different from the ratio for appropriate age students. When examining the ratio of males to females in the appropriate age group it is noted for Grades 4 and 5 the number of males and females is almost equal. In the retained group each grade has at least two times the number of males to females. This pattern is not found in the age appropriate group. Therefore, the pattern surrounding the gender of the

students who are retained is quite different from the pattern of gender observed in students placed in their correct grade.

When the two patterns observed for retained students were combined this researcher observed that retained students are usually male and scored lower than their appropriate age peers. This would indicate that males, who are retained in a grade, are less likely to be as successful as their appropriate age peers in the same grade.

## CHAPTER FIVE: DISCUSSION AND CONCLUSIONS

There are four sections within Chapter Five. The first section discusses the findings of the inter-rater reliability study. The second section considers gender, relative-age differences and the interaction between them, effect size comparisons, performance group differences and grade retention. The third section examines of the limitations of this study. The final section considers implications for educational practice.

### Inter-Rater Reliability

The inter-rater reliability study was vital to determine if the many markers of the CBM math data marked the probes reliably. Without these results, the results of the main study analyses were suspect. This pre-study was therefore vital to provide the confidence in the marking of the math probes so that the researcher could proceed with the main study.

The results of the inter-rater study indicated that marker reliability appeared to be high for the CBM math norming project. An inter-rater reliability correlation mean of .98 and median of .99 suggests that most markers reached scoring agreement. Observation of the CD scores in Table 1 confirms that most markers could agree on which probes earned high CD scores and which ones earned low CD scores. These findings are similar to inter-rater correlation means found by most other researchers. Fuchs, Fuchs, and Hamlett (cited in Marston, 1989) report an identical mean for interscorer agreement on CBM math measures. This would indicate that one could conclude high inter-rater reliability for this study.

However, from further observations beyond the mean and median of the markers it is evident the markers did not all reach agreement on the CD scores of the probes. From the  $Q_1$  and  $Q_3$  CD scores for markers and probes, it was noted that in most cases markers reached agreement within two or three CD points. Hence, if this were the only analysis used to determine inter-rater reliability one would conclude the markers were nearly identical in their marking. Thus, no cause for concern exists regarding marking of CBM math probes. However, the range in correlations indicated a few markers differed substantially in their marking from others. If the goal is to achieve a high level of inter-rater reliability further analyses of causes for marking differences indicates there are factors to consider when marking CBM math probes. Discussed below are factors which require consideration when marking CBM math probes.

### *Marking Considerations*

#### *Addition Errors*

Investigation of addition errors demonstrated the impact they can have on a probe score. Marker 19 with the largest addition error produced some of the lowest correlations and the largest marker coefficient of variation value. Correction of addition errors did not impact the correlation mean and median. However, it did decrease the range in the marker correlations by increasing the value of the minimum correlation from .78 to .83. Hence, addition errors are one factor that contributed to marker differences. Therefore, markers should take steps to ensure addition errors are not contributing to marker differences. Markers should take steps to prevent addition errors from influencing the CD score used for student evaluation and assessment. Recommendations for eliminating

addition errors include rechecking addition calculations by the marker or another marker. The person checking addition totals would not need to be competent at scoring CBM math probes only competent at adding CD scores. Increasing familiarity with the probes also assists in reducing addition errors as markers recognize an unrealistic score, for the number of questions accurately completed by a student. Elimination of addition errors therefore is vital to maintaining marker reliability.

Elimination of hand marking CBM math probes would be an alternative method of reducing addition errors. Programs which would allow machine scoring of CBM probes would assist in eliminating addition errors made by markers. Investigation into ways of addressing machine marking whether by a scanner or a CBM computer program may be worth consideration. Such a tool may have an additional advantage of reducing the amount of time spent marking CBM math probes while at the same time increasing the accuracy of scoring probes.

#### *Question Score Discrepancy*

Other issues that contributed to marker differences revolve around the discrepancy noted for specific questions. As the markers reached unanimous agreement on the CD score for only 40% of the questions, it is desirable to increase the percentage of agreement. Increasing marker agreement is one method of increasing marker reliability. Even though the probes were chosen to produce the most marker variability, it is desirable to ensure all markers reach agreement on all types of questions. Addressing the following goals should increase the percent of marker agreement on the CD score a student achieved on specific questions.

*Unmarked questions.* As was evident from analyzing the probes one cannot assume a question left unmarked earned a score of zero. Therefore, it is imperative markers check over a probe on both the front and back of the probe after marking is completed. This is to ensure every attempted question is marked. Markers cannot assume a student did not attempt questions on the back of the probe because they did not complete all of the questions on the front. Nor can markers assume students attempt questions in the order they are presented on the probe. Students do skip over questions. If a question attempted by a student earns a CD score of zero this researcher recommends the marker indicate the score earned in the question box. This will assist in reducing the possibility of leaving a question unmarked.

*Rules not followed.* Table 7 indicates rules not followed by all the markers. What is unknown is whether markers did not follow a rule because they disagreed with it, chose not to follow it, or if they forgot about the rule. No matter the reason for not implementing the rule, it caused marking discrepancies. Therefore, a recommendation for markers is that they review the marking rules before they begin marking CBM math probes. It may be necessary to review the rules more than once during the process of marking probes. Reviewing the rules would be especially important if there is a large number of probes to mark, the marker is inexperienced, a long time elapses between marking sessions, an unusual answer is given, or a question is answered that is not typically attempted by most students.

*Additional rules.* More rules may be required to cover the other issues causing marker discrepancies as noted in Table 7. For example, a specific statement to regarding



credit for positive and negative signs even if some or all of the digits are incorrect would reduce many marker differences.

*Other marking issues.* The researcher is aware there are further potential causes of marker discrepancy experienced elsewhere but not evident in the probes marked by the markers. These include the following:

- a. Rounding the answer when converting a decimal to a percent.
- b. Writing the remainder of a long division question as a decimal instead of a fraction. As decimal fractions are taught as part of the math curriculum, it might be a worthwhile consideration to provide alternate answers for questions which could have a decimal, fraction or remainder as an answer. Then students would not be penalized for completing a question correctly, using a method they have been taught, even if the answer key does not provide that alternative as a correct answer.
- c. Penalizing a student for completing long division questions the short way but missing only the decimal by giving credit for only the correct number of digits rather than deducting a mark for the missing decimal.
- d. Students receiving credit for correct digits even though they do the wrong calculation and therefore missed the concept of the question.
- e. Providing a statement addressing what should happen when students provide unnecessary numbers, figures or calculations for a question. Alternatively, a statement to ignore extra numbers because students have reduced their math fluency by spending time doing extra work, instead of completing another question.

### *Impact of Marker Differences*

If the CBM math data is to be used to assist in making educational decisions for a student then it is important to understand how marking differences between markers influences those decisions. First, it is important to recognize for which students marking differences impact the most. A student represented by Probe 1 achieved a mean score of 113 CD. For all three CBM norming periods, a score of 113 CD places this student in the 90th to 95th percentile on the CBM math norms. The most severe marker gave Probe 1 a score of 106 while the most lenient a score was 121. It is only for the spring norming period that a score of 106 decreases the percentile rating of this student to the 85th to 90th percentile. Although the student represented by Probe 1 experienced marker differences, little impact is evident as this student is always in the Above Average or Well Above Average range. It is unlikely that the educational decisions for this high performing student will change due to the variation in scores from different markers. However, the impact of marking differences is not consistent among students.

While it appears that differences in marking for high performing students may have limited effect, it is necessary to determine if this holds true for all students. Probe 12 represents a student who achieved an average score of 19 CD. This student obviously is not achieving at the same level as the student represented by Probe 1. Scores for this student ranged from a low of 15 to a high of 22. In all three CBM norming periods this range of scores impacts the percentile range in which this student is performing. For example, in the fall a score of 15 placed the student in the fifth to tenth percentile. This is the Well Below Average range. In contrast, the high score of 22 moved the student up to the 10th to 15th percentile range, the Below Average range. If this CBM math score is

used to set educational goals, or to request extra assistance for the student, the more severe score would increase the students chances of receiving support, as it places the student within the Well Below Average range.

The impact of marker differences therefore, is not consistent for all students. While students who perform in the Below Average or Way Below Average range may not always demonstrate variation in their scores, they may be the ones to be impacted the most by marker differences. It is important therefore, that markers recognize which students may be impacted by marking differences and take steps to eliminate any causes of marker differences.

#### *Recommendations to Reduce Marker Differences*

If this assessment tool is to be reliably used by educators to make educational assessments, decisions and comparisons providing and maintaining consistent, reliable markers for CBM math is a worthwhile goal.

One suggestion to maintain marker reliability is to provide further inservice on marking the CBM math probes. Until all markers are comfortable and familiar with the process of marking, inservice and training may be required on a regular basis. Regular training will also ensure staff members who did not receive previous CBM math administration and marking inservice also have the opportunity to receive training. Markers could use training inservice as a way of sharing questions with unusual answers. Thus, it would be an opportunity to reach marking consensus for questions with unusual answers.

Another way to increase marker reliability is to have markers work together when scoring the math probes. Two educators from each school were invited to attend the CBM math training inservice. It is possible that in several schools two trained markers work in the same building. Thus, it is possible for two markers to work together on CBM marking. The experience of marking with another person would provide markers with the opportunity to discuss options for marking challenging questions. Even if unable to get together with someone else to do the marking, markers are encouraged to ask another educator who is more familiar with the process. Thus, by asking questions a marker is marking with someone else without the proximity of the person. By marking with other people in person or by communicating with someone else, it is possible to verify scores and answers. Hence, marker agreement could be achieved between the markers working together.

The designers of the CBM probes may consider development of further CBM math rules to answer some of the issues, which caused marking discrepancies. In some cases, additional examples or statements might provide the information necessary to assist markers in achieving agreement about how to mark a specific question.

Implementation of any of these suggestions would assist in maintaining a high level of marker reliability. A minimum goal should be to maintain the level of inter-rater reliability found in this study. Preferably, educators would strive to increase the inter-rater reliability. More importantly, the goal would be to increase the percentage of marker agreement on questions beyond the present 40%. Several options are available to accomplish this goal. Thus, maintaining consistent marking would ensure the educational

decisions made for students are not impacted by the marking differences of a lenient or severe marker.

While there is evidence to conclude that inter-rater reliability of CBM math probes is high, the conclusion is not without a few considerations. Presently, there are several factors reducing the agreement between markers. While any one of these factors might play only a small role in reducing marker agreement, their cumulative effect is evident as demonstrated by the range of inter-rater correlations (.78 to 1.00). These are recommendations, if implemented should increase agreement between markers. While the inter-rater study indicates there is a high degree of agreement between markers, it would not be realistic to accept that no improvement is possible or necessary.

### The Main Study

Within the main study, discussion encompasses four topics. The first topic considers gender and relative-age differences. Next, discussion looks at effect size comparisons to compare the results of gender differences. The third topic looks at the existence of performance group differences. Finally, the discussion considers grade retention.

Before discussing the results of the main study, it is important to remember the CBM math probes used for this study focus on math computation. CD scores are calculated from the number of correct digits a student computes in five minutes. As with other CBM measures the math probes follow a set of procedures created by the developers of CBM (Fuchs & Fuchs, 1992). The CBM math probes do not test all aspects of the elementary mathematics curriculum. Some topics the CBM math probes do not

cover include statistics and probability, and shape and space. Nor do the CBM math probes investigate how a student applies their mathematical knowledge to higher level problem-solving. Information regarding the elementary math curriculum is available in the Mathematics K to 7 Integrated Resource Packages (Province of British Columbia, 1995). Educators cannot expect to use the probes to test the complete range of the math curriculum but can expect to see a small picture of a student's math knowledge. What is presently unknown is how a student's rating or percentile score on a CBM math probe correlates with their knowledge in other parts of the math curriculum.

### *Gender and Relative Age Differences*

Ongoing concern has surrounded the issues of gender and relative-age differences. Other researchers have not reached agreement about the existence of these issues. Discussion surrounding the main study will consider gender differences, relative-age differences and the interaction between gender and relative-age.

*Gender differences.* The main analyses for gender differences, using a repeated-measures ANOVA determined found no evidence of their existence in any grade. However, further investigation by calculating effect size differences conflicted with the results of the 2 x 3 x 3 between-subjects repeated-measures ANOVA for Grade 2 and the 2 x 3 ANOVA for Grade 1. Calculation of effect sizes provided evidence that a consistent pattern of noticeable gender differences were evident throughout the Grade 1 and all of the Grade 2 data. This conflicted with the nonsignificant main analysis. However, with the exception of the January and April Grade 7 data no evidence of gender differences was evident in other grades. This conflict prompted further research to determine which analysis was accurate. The *t* test this researcher determined there were gender differences

in the Grade 1 data and the October and January Grade 2 data. According to the  $t$  test, no gender differences existed in other grades. Effect size and  $t$  test results indicated that gender differences did exist for Grade 1 and Grade 2. The results for these two grades remained in conflict with the main analysis. The Box M test producing significant results  $p < .001$  at Grade 2 may have been an indication that the analysis for this grade was not accurate. The  $t$  test, effect sizes, and repeated-measures ANOVA provided evidence that no gender differences existed for Grades 3 to 6. Effect sizes showed a small gender difference for January and April norming periods of Grade 7. In contrast, the main analysis and  $t$  test both indicated no Grade 7 gender differences. From the direction calculated for each effect size it is evident that one gender is not consistently favoured over another in math. However, the gender differences noted for Grades 1 and 2 from the  $t$  test and effect sizes consistently favour boys. Thus, this researcher concluded that gender differences in favour of boys exist only in Grades 1 and 2 but gender differences do not exist within Grades 3 to 7.

Several aspects of this study appear to contradict other research. Existence of significant gender differences at the early primary grades contrasts with the findings of other studies. Leahey and Guo (2001) determined that elementary students have equal starting points. This is not what the evidence from this study shows. On the other hand, Cole (1997) found a slight increase in gender differences from Grades 4 to 8. Again, this contrasts with the findings of this study, which found differences in the early primary grades but not Grades 3 to 7. Many researchers suggested the concerns surrounding gender differences exist as students reach secondary school. This study did not investigate students in secondary grades so it is not possible to confirm the results found

for secondary students but trends in the data do not indicate an increase in gender differences as students reach higher grades. Another contradiction is the research by Hay et al. (1998) and Beaton et al. (1996) who found gender differences favour girls. However, this researcher did not find either gender consistently favoured. In addition, with the exception of Grades 1 and 2 the effect sizes are trivial and therefore the size and direction is not important. Thus, several contradictory results exist between this research and that of other studies.

For SD57 the results of this present study for CBM math are different from the gender differences discovered by Hedekar (1997) and MacMillan (2000) for CBM reading and writing fluency. They found consistent gender differences at all grades, which disagrees with this study. This study only found gender differences at Grades 1 and 2. Whereas in this study the gender differences for Grades 1 and 2 favour boys the results is that Hedekar (1997) and MacMillan (2000) found girls were favoured in their study in all grades. However, without further investigation, it is impossible to conclude whether the differences are due to the subject matter or because a change in gender differences has taken place in the same school district within a few years.

For Grades 3 to 7, this study confirms the results of research which did not find evidence of gender differences. Ma's (1999) results show no gender differences in Grade 7. Willingham and Cole (1997) also agreed with the findings that gender differences in math do not exist. Despite contradiction with some research, there is agreement among some research and the results of this study.

This study is prolonging the debate surrounding the existence of gender differences. Finding evidence of gender differences at Grades 1 and 2 is not supported by



other researchers. However, the confirmation by effect sizes and  $t$  test indicate that it exists. The evidence from this study suggests that concern on gender differences at the higher grades is misplaced, and educators should be concerned about the gender differences that favour boys in the early primary grades. Hence, for this study it can be concluded that gender differences in math do exist in Grades 1 and 2. These results indicate that in early primary grades males will outperform females in math achievement. In contrast, the evidence from the main analyses, effect sizes, lack of consistent effect direction and  $t$  test verify the conclusion that gender differences do not exist at Grades 3 to 7. Even if they previously existed for Grades 3 to 7, they are no longer evident. Presently therefore, neither boys nor girls in Grades 3 to 7 have an academic advantage in math.

*Relative-age differences.* This study did not find relative-age differences in math. Within the main analysis, Grade 3 produced the only significant relative-age difference. All other grades produced nonsignificant results. This indicated no consistent relative-age difference trend or pattern. Further analysis of these results using effect sizes confirmed a lack of consistent results with most effects less than small. The few small effect size results were not consistent within a grade, relative-age, or norming period. Nor did the small effect size results show evidence of a pattern. This lack of relative-age differences was further confirmed by the  $t$  test, which did not find a consistent significant difference within a grade, relative-age, norming period or evidence of a pattern. Hence, the result of this study fail to demonstrate consistent relative-age differences exists.

These results confirm research indicating a lack of relative-age effect influencing academic achievement. This research agrees with the findings of Gredler (1992) and Bickel et al. (1991) that relative-age differences are not an issue.

Hedekar (1997) and MacMillan (2000), using reading and writing data from the same school district also found no evidence of gender differences. For SD57 this is good news. Relative-age differences were not evident for reading and writing, nor are they evident for math. This research therefore is not isolated in its findings.

This study disagrees with the findings of other researchers. Gullo and Burton (1992) found relative-age was one factor predicting academic achievement for pre-first-grade students. This study does not agree with Bisanz et al. (1995) who found relative-age does influence conservation of number for Kindergarten and Grade 1 students. Nor do the results agree with the research on sport achievement. Glamser and Marciani (1992) and Boucher and Mutimer (1994) discovered relative-age impacts achievement in sports up into adulthood. In comparison to sports, relative-age is not affecting academic achievement at any elementary age. The research in the field of science by Bell and Daniels (1990) also contradicts the results of this study on relative-age. For much research, the results showed relative-age does play a role in achievement.

The lack of relative-age difference adds to the research stating it has no effect on math achievement. These findings add to the controversy surrounding the influence of relative-age on achievement. However, for present time these results demonstrate relative-age is not a factor influencing math achievement.

*Gender and relative-age interactions.* Initially the evidence indicated only one significant interaction in Grade 7 was evident in the data. However, the  $p$  value of .11

from the Grade 1 results made it a value of interest. This factor, along with the significant Box M result for Grade 2, prompted further investigation. This investigation showed an interaction is evident at Grade 1 and 2 between gender and relative-age. From the interaction graphs for Grade 1 students it was evident that the oldest students were most impacted the interaction with gender. In contrast, the young students in Grade 2 were impacted by the interaction with gender. This reversal of which relative-age is impacted most by the gender interaction indicates that a consistent effect is not evident. Observation of the effect sizes confirms this lack of pattern not only for Grade 1 and 2 but also for Grade 7 students. Therefore, the lack of evidence of consistent interaction or pattern indicates that a gender and relative-age interaction is not evident.

This lack of gender and relative-age interaction is not new to other researchers. Hedekar (1997) also noted a lack of gender and relative-age interaction in the CBM reading and writing data from SD57. Gullo and Burton (1992) also did not find an interaction for relative-age and gender. Hence, research agrees about the lack of interaction between gender and relative-age. Educators therefore do not need to be concerned that the combination of gender and relative-age of a student will impact the student's math achievement.

*Effect size comparisons.* Comparison of effect sizes indicated the existence of similarities and differences between the results of this study and that of other researchers. The first effect size comparison discussed will be that of CBM reading and writing. Then effect size comparison with other math research is considered.

It was evident substantial differences exist when the results of this study were compared with the CBM reading and writing (WSC) results from Hedekar and

MacMillan (personal communication, January 30, 2002). As described in Chapter Four, CBM math results do not consistently favour one gender over the other. In addition, most CBM math results are trivial in size. In comparison, all CBM reading effects favour girls. With two exceptions the magnitude of the CBM reading results are larger than the CBM math results. As with CBM reading, the writing results consistently favour girls. Except for Grade 1, all CBM writing results are larger in magnitude than for CBM math. In fact, some CBM writing results reach a medium size result. This is a much larger effect than the trivial CBM math effects. Effect sizes are therefore significant and consistent in CBM reading and writing but are not significant in CBM math. The effect sizes results are a further indication that consistent gender differences were not apparent in the CBM math results but were in the CBM reading and writing. Hence, gender differences in math are not a concern.

Comparison of the CBM math effect size results with other researchers provided information regarding how different the results of this study are with other findings. Even though the ongoing gender debate indicates cause for concern regarding gender differences, comparison of the effect sizes provides another viewpoint. Three of the effect sizes chosen for comparison from this study show a small effect size magnitude. All others were trivial in magnitude. The effect sizes of all other researchers reported in this study were trivial in magnitude. All but two researchers showed a trivial effect favouring boys. Comparison of the results of this study and other researchers leads one to conclude that at Grade 4 the magnitude is identical. At Grade 7 the effect size of this study is slightly larger but not by a substantial amount than all other researchers. Therefore, it is possible to consider the effect sizes at Grade 7 of this study as similar to those of other

researchers. A similarity between this study and other researchers is the inconsistent pattern favouring one gender over the other. While some effect size results of this study might be slightly larger than other researchers, they do not appear to be substantially different. Therefore, the effect sizes of this research are similar to the results found in other math studies.

### *Performance Group Differences*

Performance group investigation found a lack of significant gender, relative-age or interaction differences for both high and low performing students. Changing the definition of *high performing* from above the 75th percentile to above the 90th percentile did not change the results. The same was true for the *low performing* group whether defined as below the 25th percentile or below the 10th percentile. These results appear to disagree with several other researchers. Fan (1995), Royer et al. (1999), and Kleinfeld (1998b) all found evidence of gender differences among high performing students. Kleinfeld (1998a) reviewed a variety of measures and found gender differences in the top 10% of the students. This researcher, on the other hand, did not find differences among the top performing group at either above the 75th percentile or the 90th percentile.

A further difference between the results of this study and other research is the ratio of males to females found in the high performing group. Willingham and Cole (1997) found a ratio of .70 females to males. This differs from the Grade 7 ratio of 1.00 calculated by this researcher. Hedges and Nowell's (1995) ratio of 1.64 males to females for the top five percent also contradicts the findings of this present study. This present

study suggests educators need not be concerned that fewer girls than boys are among the high achievers.

The findings of this research confirm those of Mullis, Martin, Fierros, et al. (2000) which indicated a significant difference did not exist between students in the top 25% of students. Because evidence of a gender difference between high performing students does not appear to exist, educators do not need to implement techniques to try to close the gap between the genders. Therefore, differences in performance groups are nonexistent.

Investigation of the ratio of students in the low performance group found the results were not in agreement with other research. Kleinfeld (1998b) noted three to four times more boys have learning disabilities. Consequently, more boys are at the bottom ability group. The results of this present study did not find a consistent pattern or trend in the ratio of students in the low performance group. Hence, this research indicates gender differences do not impact math achievement in the low performance group.

To determine if there is a further reason to be concerned about high performing students examination of relative-age differences in high performing students provided the opportunity. This study did not find a significant relative-age effect for performance groups. Nor did it find an interaction for gender and relative-age for any of the performance groups. Therefore, Sweeney's (1995) results contradict the findings of this study. It is not necessary for parents and educators to be concerned that the gender, relative-age and performance group determines the success of a student. Hence, educators do not need to address gender and relative-age differences in performance groups.

### *Grade Retention*

Most retained students did not do as well in a grade below what was appropriate for their age, as did the students in the same grade who were the correct age for the grade. Although the reasons for retaining students is unknown, the strategy of retention is not providing students with the opportunity to obtain the same scores academically as their peers who are in the appropriate grade for their age. More boys than girls were in the retained group. The results of this study agree with other researchers that retention has a negative impact on students (Foster, 1993; Owings & Magliaro, 1998; Shepard & Smith, 1987). The lower mean CBM math scores obtained by retained students demonstrates that retained students have lower performance than their age appropriate peers in the same grade. Grade retention is a strategy that does not guarantee student success. This study demonstrates that male students, struggling academically in the grade that is appropriate for their age do not become a high or even average academic achiever when retained in a grade below what is appropriate for their age.

The results of this study suggest grade retention fails to address ways to assist students who are struggling academically. Hence, parents and educators should not recommend grade retention as a solution for students who are not experiencing academic success.

### *Limitations of the Study*

There are several limitations influencing the results of this study. Discussion regarding the limitations is in two sections. First, the limitations of the inter-rater study will be addressed. Then, the researcher examines the limitations of the main study.

### *Inter-Rater Study*

Several limitations existed in the inter-rater study. In order to provide all markers with the same probes they experienced the disadvantage of marking photocopies rather than original papers. Therefore, it was not possible to investigate factors affected by marking original work. In addition, two factors limited the selection of probes the researcher had to choose from for the inter-rater study. First, not all schools had sent in the probes of the students who participated in the CBM math norming project. As a result, the selection of probes was limited to the probes received. Secondly, probe selection was further limited due to the marking techniques used by some markers. Probes were eliminated from being selected for the study if the marker put marks over the student answers in a manner that made it impossible for the researcher to white them out and photocopy the probe while maintaining the integrity of the students' answers. The inter-rater study provided a wide selection of marking differences to examine. However, due to limitations of the study there were factors the researcher did not investigate. The researcher did not investigate the following issues for the inter-rater study:

1. Marking of erased or faintly written answers. Due to the need to photocopy marker packages from the student norming probes, original probes were not marked by the markers and it was necessary to ensure all probes used would consistently be readable when copied.
2. Rule 6 regarding reversed digits was not investigated as no reverse digits were evident in the probes chosen. This would be more apparent in probes completed by early primary students.



3. Rule 7 regarding rotated digits was also not investigated. As with Rule 6 there were no rotated digits on the probes.
4. Several of the issues causing marker discrepancies do not generalize to markers of all CBM math probes. Probes for many of the earlier grades do not possess the type of questions, which produced marker discrepancies. Conversely, if potential issues causing marker discrepancies for younger grades exist, this study was not aware of them.
5. Marking a sampling of the six different Grade 7 probes at one time opposed to marking a multiple number of the identical probe may have contributed to marker discrepancies. When marking the same probe several times a marker becomes familiar with the challenging questions to look for, and the typical number of questions completed by a high or low performing student. Markers can also compare scores and the questions answered when marking multiples of an identical probe. It is possible therefore, that marking package containing a sampling of all six norming math probes reduced the familiarity markers develop which assists in decreasing marking differences.

Although, many limitations existed in the inter-rater study it provided valuable information regarding marking differences between markers of the CBM math probes. Limitations, which remain a concern for markers and developers of CBM math probes, could be investigated at another time.

### *The Main Study*

Within the main study several limitations are apparent. A limitation may have contributed to the lack of gender differences in the performance groups. As the CBM math probes test math computation only, they do not test students' ability to apply the calculations to problem-solving activities. Whereas average or low performing students often struggle with problem-solving and application questions these are often the type of questions at which high performing or high ability students excel. Without high-level application questions on the CBM math probes to provide this technique to separate the high performing students from other students, the performance group analysis may have been limited in the ability to successfully identify high performing students.

Comparison of the results obtained with the CBM math probes and other achievement tests has yet to be undertaken. Therefore, because only one testing instrument is used, the evidence produced from the study may be limited to computation tasks. Comparison of other test results to the CBM math study, with the same students would indicate whether the results of this study are limited to this specific testing resource or generalize to all types of testing resources.

A further limitation of this study revolves around the students selected to participate in the study. All transient students were removed from the CBM norming sample for this research. If transient students represent a sub-group of students with specific characteristics different from the rest of the population, they are not represented in the larger population chosen to participate in this study. Thus, it is unknown at this time if the results of this study generalize to transient students or if they represent a population with different characteristics.

Although, limitations are evident in this study it remained a worthwhile endeavour to determine gender and relative-age differences in math, which may exist between students in SD57. In addition, the insight provided by the inter-rater study into the marking of the CBM math probes reduces suspicion regarding their potential usefulness for measuring and comparing student progress.

### Implications for Future Study and Practice

This section considers the implications of this study as it relates to inter-rater reliability, gender and relative-age differences, performance groups, and grade retention.

The inter-rater correlation mean of .98 for the CBM math norms confirmed a high degree of reliability for the markers of the probes. This indicates that it is possible to train a large number of teachers to accurately administer and mark these assessment tools. The findings of the inter-rater study indicate educators can confidently utilize the CBM math norms. Results of the CBM math norms are no longer suspect to unreliable marking practices. Thus, educators can explore further ways to implement the CBM math norms as part of their everyday educational practice and as an alternate assessment tool. However, educators are reminded that the probes created for the CBM norms do not cover all aspects of the mathematics curriculum. In SD57 effective mathematics assessment covering all aspects of the curriculum would require additional assessment tools beyond the CBM math norming probes.

Further inservice addressing the limitations of the inter-rater study would serve to increase the marker reliability. This would ensure CBM measures used to assess and monitor student progress are not influenced by marker differences. Future research could

determine three issues regarding inter-rater reliability. The first concern would examine how additional inservice changes marker reliability for educators who received the initial training. Another issue to investigate is the reliability of markers, presently using CBM measures that did not attend the school district inservice. A third question to examine is how the reliability of trained markers compares to markers with informal or no training. As the use of CBM math measures increases among educators, these future research questions are worth investigating to maintain a high degree of marker reliability.

There is one population not investigated within the analyses of the main study. Elimination of students who were missing at least one CD score were eliminated from the norming sample for the analyses of the main study. Either these were transient students or students who were absent during the testing time. Transient students are students who move at least once during a school year. These students did not make up part of the sample for the main study. A total of 181 students were eliminated for missing data. However, not enough students were present in each grade and norming period to guarantee a robust analysis (Tabachnick & Fidell, 2001) as 11 out of the 18 cells would have fewer than 20 students. In addition, information regarding a student's absence was unavailable to determine if the missing data was due changing schools or another reason. Therefore, it was not realistic to perform this analysis. A direction for future research might investigate how these students compare with students who remain at the same school for the whole year. Presently, it is unknown if and how transient students compare to the rest of the student population. Without researching this information, it is difficult to determine if transient students require a different focus or assistance from the other students.

Educators in SD57 have now had access to the CBM math norms for almost two years. Many aspects regarding CBM math norms remain unidentified. Their implementation and use in each school is unknown. Nor has their concurrent and predictive validity been determined. SD57 has not undertaken research to determine how they compare to other math assessments, standardized tests, teacher, school, or district created tests. Researching these issues would provide evidence regarding the usefulness and validity of CBM measures.

Results from the main study provide educators with confidence that gender differences between students are not a concern in Grades 3 to 7. Findings of main analysis demonstrated gender differences exist between males and females in Grades 1 and 2 in math achievement. Although the results from subsequent grades indicate females catch-up, the cause of this difference was not determined. Gender differences evident in Grades 1 and 2 in favour of boys will need to be addressed. In addition, it may be worthwhile to investigate interventions that assist females and males to equally reach their academic math potential in Grades 1 and 2. A future focus could investigate causes of the difference. Alternatively, since girls catch-up to boys in later grades, ignoring this difference does not predict future failure.

Relative-age differences often thought to impact younger students in the same grade are not an issue requiring intervention. No advantage was found to being one of the older, average or younger students in a grade. Consequently, educators do not need to pursue interventions to address this issue.

Although this study did not find performance group differences, this issue cannot be forgotten. Investigation of this concern should continue utilizing assessment tools

requiring high-level problem-solving skills. CBM techniques could be considered if high-level problem-solving and application skills can be utilized as part of the assessment. Then if further research indicates no gender or relative-age differences are evident in math achievement, concern regarding performance group differences will not be necessary. According to the results of this study, one gender is not outperforming another in any of the performance groups. Thus, interventions are presently not required to assist one gender over the other in one or more of the performance groups.

As this study indicates, grade retention is not a successful technique to guarantee academic success for students. The issue of grade retention requires examination to establish why grade retention is considered for students. In addition, investigation could determine why more boys experience grade retention more than do girls. As grade retention does not lead to successful math achievement alternatives require exploration.

This study attempted to answer questions regarding math achievement for elementary students. Research might now determine if secondary students replicate the results found for elementary students within the same school district.

While providing answers to questions regarding CBM math this study has also created more questions for consideration. Attempts to answer any of these questions will provide educators with further information regarding the implementation and use of CBM math and the state of math differences between students.

For the moment, CBM math norms can provide educators with the confidence that few differences exist between students according to gender, relative-age and performance groups. The differences that exist are minimal. Gender differences for

Grades 1 and 2 soon disappear. The grade retention results were anticipated and reconfirm what research has previously indicated.

## REFERENCES

- Allinder, R. M. (2000). Effects of teacher self-monitoring on implementation of curriculum-based measurement and mathematics computation achievement of students with disabilities. *Remedial & Special Education, 21* (4). Retrieved July 31, 2001, from EBSCOhost: <http://ehostvgw5.epnet.com/fulltext.asp>
- Allinder, R. M., & Eccarius, M. A. (1999). Exploring the technical adequacy of curriculum-based measurement in reading for children who use manually coded English. *Exceptional Children, 65* (2). Retrieved July 10, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- American Association of University Women Educational Foundation. (1992). *How schools shortchange girls. A study of major findings on girls and education. The AAUW report.* Wellesley College: MA. Center for Research on Women. (ERIC Document Reproduction Service No. ED 339 674)
- Baker, S., Collins, V., & Goodwin, M. (1992). Administration and scoring of curriculum-based measurement. In M. R. Shinn, N. Knutson, & W. D. Tilly III (Eds.), *CBA training institute 1992.* University of Oregon: Oregon.
- Barnsley, R. H. (1988). *Birthdate and performance: The relative age effect.* Paper presented at the Canadian Society for the Study of Education. Windsor: ON. (ERIC Document Reproduction Service No. ED 306 679)
- Beal, C. R. (1999). Special issue on the math-fact retrieval hypothesis. *Contemporary Educational Psychology 24*, 171-180.
- Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College. Retrieved January 7, 2001, from the World Wide Web: <http://timss.bc.edu/timss1995i/TIMSSPublications.html>
- Bell, J. F., & Daniels, S. (1990). Are summer-born children disadvantaged? The birthdate effect in education. *Oxford Review of Education, 16* (1). Retrieved July 31, 2001, from EBSCO host: <http://ehostvgw.wysiwyg:bodyframe.main.3/>
- Beller, M., & Gafni, N. (1996). The 1991 international assessment of educational progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology, 88* (22), 365-377.



- Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, 42 (1/2), 1-21. Retrieved February 10, 2002, from Pro Quest: <http://virtue.unbc.ca:2107/qudweb?>
- Bickel, D. D., Zigmond, N., & Strayhorn, J. (1991). Chronological age at entrance to first grade: Effects on elementary school success. *Early Childhood Research Quarterly*, 6, 105-117.
- Bisanz, J., Morrison, F. J., & Dunn M. (1995). Effects of age and schooling on the acquisition of elementary quantitative skills. *Developmental Psychology* 31, 221-236.
- Boucher, J. L., & Mutimer, B. T. P. (1994). The relative age phenomenon in sport: A replication and extension with ice-hockey. *Research Quarterly for Exercise and Sport*, 65 (4). Retrieved August 1, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- Boyd, P. C. (1989). *The relationship to age entrance to kindergarten to achievement on grades one through five*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association. Reports - Research. (ERIC Document Reproduction Service No. ED 313 156)
- Canadian students near top of the class in math, science; International test results. (2000, December 6). *National Post Online*. Retrieved December 9, 2000 from the World Wide Web: [www.nationalpost.com/search/story.html](http://www.nationalpost.com/search/story.html)
- Cohen, J. (1992). A primer power. *Psychological Bulletin*, 112, 155-159.
- Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings*. Princeton, NJ: Educational Testing Service. Reports – Research. (ERIC Document Reproduction Service No. ED 424 337)
- Crosser, S. L. (1991). Summer birth date children: Kindergarten entrance age and academic achievement. *Journal of Educational Research*, 84 (3), 140-146.
- Daniels, V. I. (1999). The assessment maze: Making instructional decisions about alternative assessments for students with disabilities. *Preventing School Failure, Summer*. Retrieved March 16, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52 (3), 219-232.

- Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1-17). New York: Guilford Press.
- Deno, S. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure*, 36 (2). Retrieved July 31, 2001, from EBSCOhost: <http://ehostvgw5.epnet.com/fulltext.asp>
- Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education*, 34 (3). Retrieved July 10, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- Fan, X. (1995). *Change in mathematics proficiency for male and female students from 8<sup>th</sup> to 12<sup>th</sup> grade: A study based on a national longitudinal sample*. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA. (ERIC Document Reproduction Service No. ED 387 348)
- Fewster, S. (2000). *School-based evidence for the validity of curriculum-based measurement norms in School District No. 57*. Unpublished master's thesis, University of Northern British Columbia, Prince George, BC, Canada.
- Foster, J. E. (1993). Reviews of research: Retaining children in grade. *Childhood Education*, 70 (1), 38-43.
- Fuchs, L. S. & Deno, S. L. (September, 1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children*. Retrieved March 16, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- Fuchs, L. S. & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, 21 (1), 45-58.
- Fuchs, L. S. & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children*, 30 (3), 1-16.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review*, 19 (1), 6-22.
- Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham, Great Britain: Open University Press.
- Glamser, F. D., & Marciani, L. M. (1992). The birthdate effect and college athletic participation: Some comparisons. *Journal of Sport Behaviour*, 15 (3). Retrieved July 31, 2001, from EBSCO host: [wysiwyg:bodyframe.main.3/http://ehostvgw5.epnet.com/fulltext.asp](http://ehostvgw5.epnet.com/fulltext.asp)

- Glass, G. V. & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Boston: Allyn and Bacon.
- Gredler, G. R. (1992). *School readiness: Assessment and educational issues*. Brandon, VT: Clinical Psychology Publishing Co., Inc.
- Gullo, D. F. & Burton, C. B. (1992). Age of entry, preschool experience and sex as antecedents of academic readiness in kindergarten. *Early Childhood Research Quarterly*, 7, 175-186.
- Hall, C. W., Davis, N. B., Bolen, L. M. & Chia, R. (1999). Gender and racial differences in mathematical performance. *The Journal of Social Psychology*, Retrieved January 21, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- Hay, I., Ashman, A. F., & van Kraayenoord, C. E. (1998). The influence of gender, academic achievement and non-school factors upon pre-adolescent self-concept. *Educational Psychology*, 18, 461-470. Retrieved January 4, 2002, from Pro Quest: [wysiwyg://16http://virtue.unbc.ca:2107/p](http://wysiwyg://16http://virtue.unbc.ca:2107/p)
- Hedekar, L. (1997). *The effects of month of birth and gender on elementary reading and writing fluency scores using curriculum-based measurement*. Unpublished master's thesis, University of Northern British Columbia, Prince George, BC, Canada.
- Hedges, L. V. & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269, 41-45.
- Howell, K. W., Fox, S. L., & Morehead, M. K. (1993). *Curriculum-based evaluation: Teaching and decision making*. Pacific Grove, CA: Brooks/Cole Publishing Company.
- King-Sears, M. E., Burgess, M., & Lawson, T. L. (1999). Applying in inclusive settings: Curriculum-based assessment. *Teaching Exceptional Children*, Retrieved March 16, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- King-Sears, M. E., Cummings, C. S., & Hullahen, S. P. (1994). *Curriculum-based assessment in special education*. San Diego, CA: Singular Publishing Group, Inc.
- Kirk, R. E. (1990). *Statistics: An introduction*. Fort Worth, TX: Holt, Rinehart and Winston, Inc.
- Kleinfeld, J. (1998a). *The myth that schools shortchange girls: Social science in the service of deception*. Washington, DC: Women's Freedom Network. Reports - Descriptive. (ERIC Document Reproduction Service No. ED 423 210)

- Kleinfeld, J. (1998b). Why smart people believe that schools shortchange girls: What you see when you live in a tail. *Gender Issues*, 16 (1/2), 47-63.
- Kranzler, J. H., Miller, M. D., Jordan L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly*, Retrieved February 8, 2001, from Pro Quest: <http://proquest.umi.com/pqdweb>
- Leahey, E., & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces*, 80 (2). Retrieved January 5, 2002, from Pro Quest: <http://www.virtue.unbc.ca:2107/p>
- Linacre, J. M., & Wright, B. D. (1996). *A user's guide to Bigsteps: Rasch-model computer program*. Chicago: MESA Press.
- Lindgren, A. (2000, December 9). Half of Ontario pupils miss mark. *The Ottawa Citizen*. Retrieved December 9, 2000 from the World Wide Web: <http://www.canada.com>
- Ma, X. (1999). Gender differences in growth in mathematical skills during secondary grades: A growth model analysis. *Alberta Journal of Educational Research*, XLV (4), 448-466.
- MacMillan, P. (2000). Simultaneous measurement of reading growth, gender, and relative-age effects: Many-faceted rasch applied to CBM reading scores. *Journal of Applied Measurement*, 1, 393-408.
- MacMillan, P. D. (2001). *Simultaneous measurement of mathematics growth, gender, and relative-age effects: Many-faceted rasch applied to CBM scores*. Paper presented at the annual conference of the Canadian Society for the Study of Education (CSSE), Laval, QU, Canada.
- Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford Press.
- Marston, D. & Magnusson, D. (1988). Curriculum based measurement: District-level implementation. In J. L. Graden, J. E. Zins, & M. J. Curtis (Eds.) *Alternative educational delivery systems: Enhancing instructional options for all students*. (pp. 137-172). Washington, DC: National Association of School Psychologists.
- Meisels, S. J. (1992). Doing harm by doing good: Iatrogenic effects of early childhood. *Early Childhood Research Quarterly*, 7, 155-174.

- Meisels, S. J., & Liaw, F. R. (1993). Failure in grade: Do retained students catch-up? *Journal of Educational Research*, 87 (2), 69-77.
- Morrow, D. & Goertzen, S. (1986). *A commentary on gender differences*. Manitoba Department of Education, Winnipeg. Planning and Research Branch. (ERIC Document Reproduction Service No. ED 301 469)
- Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the primary school years: IEA's third international mathematics and science study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College. . Retrieved January 7, 2001, from the World Wide Web: <http://timss/bc.edu/timss1995l/timsspdf/amtimss.pdf>
- Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. E. (2000). *Gender differences in achievement: IEA's third international mathematics and science study (TIMSS)*. Chestnut Hill, MA : The International Study Center, Lynch School of Education, Boston College. Retrieved January 9, 2002, from the World Wide Web: <http://timss.org/timss1995i/gender.html>
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (1999). *TIMSS 1999 international mathematics report*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College. Retrieved January 7, 2001, from the World Wide Web: [http://timss.bc.edu/timss1999i/math\\_achievement\\_report.html](http://timss.bc.edu/timss1999i/math_achievement_report.html)
- Narahara, M. (1998a). *The effects of school entry age and gender on reading and math achievement scores of second grade students*. Reports - Research. (ERIC Document Reproduction Service No. ED 421 233)
- Narahara, M. (1998b). *Kindergarten entrance age and academic achievement*. Information Analyses. (ERIC Document Reproduction Service No. ED 421 218)
- Olson, G. H. (1989). *Date of birth and its effect upon performance in school over subsequent years*. Paper presented at the Annual Meeting of the American Educational Research Association. San Fransisco, CA. Reports - Research. (ERIC Document Reproduction Service No. ED 307 289)
- Owens, A. M. (2001, March 1). Boys should start kindergarten a year later than girls, report advises. *National Post Online*. Retrieved March 16, 2001 from the World Wide Web: [www.nationalpost.com/search/story.html](http://www.nationalpost.com/search/story.html)
- Owings, W.A. & Magliaro, S. (1998). Grade retention: A history of failure. *Educational Leadership*, September, 86-88.

- Province of British Columbia (1995). *Mathematics K to 7: Integrated resource package 1995*. Victoria, BC, Canada: Ministry of Education, Curriculum Branch.
- Rabinowitz, L. G. (1989). *School entry age: The effects on school achievement and adjustment. An education field problem research project report*. Reports - Research. (ERIC Document Reproduction Service No. ED 307 041)
- Royer, J. M., Tronsky, L. N., Chan, Y., Jackson, S. J., & Marchant, III, H. (1999). Math-fact retrieval as the cognitive mechanism underlying gender differences in math test performance. *Contemporary Educational Psychology*, 24, 181-266.
- Sadker, D. (April, 1999). Gender equity: Still knocking at the classroom door. *Educational Leadership*, 22-25.
- Salvia, J. & Hughes, C. (1990). *Curriculum-based assessment: Testing what is taught*. New York: MacMillan Publishing Company.
- Sax, G. & Newton, J. W. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont, CA: Wadsworth.
- School District No. 57. (1996). *Guidebook for the use of curriculum based measurement in School District #57*. Prince George, BC, Canada: School District No. 57.
- School District No. 57. (2000). *Draft norms tables for CBM math calculation*. Prince George, BC, Canada: School District No. 57.
- Shepard, L. A., & Smith, M. L. (1987). Effects of kindergarten retention at the end of the first grade. *Psychology in the Schools*, 24, 346-357.
- Shinn, M. R. (1989). Preface and Acknowledgments. In M. R. Shinn (Ed.), *Curriculum-based measurement*. (pp. v-viii). New York: Guilford Press.
- Shinn, M. R., Nolet, V., & Knutson, N. (1990). Best practices in curriculum-based measurement. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-II* (pp. 287-307). Washington, DC: The National Association of School Psychologists.
- Sweeney, N. S. (1995). The age position effect: School entrance age, giftedness, and underachievement. *Journal for the Education of the Gifted*, 18, 171-188.
- Tabachnick, B. G. & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.
- Tucker, J. A. (1985). Curriculum-based assessment: An introduction. *Exceptional Children* 52, 199-204.

- Walraven, G. & MacMillan, P. (2000). *Draft technical report of the curriculum based measurement (math) norming project*. Unpublished report, University of Northern British Columbia, Prince George, BC, Canada.
- Warder, K. (1999). *Born in December: Ready for school?* Reports - Research. (ERIC Document Reproduction Service No. ED 439 815)
- Wentzel, K.R. (1988). Gender differences in math and english achievement: A longitudinal study. *Sex Roles: A Journal of Research*, 18, 691-99.
- Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

## APPENDICES



## Appendix A

### Letters of Permission

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

3333 University Way, Prince George, BC V2N 4Z9

Dr. Alex Michalos  
Chair, UNBC Ethics Review Committee  
Tel: (250) 960-6697 or 960-5011  
Fax: (250) 960-5746  
E-mail: [michalos@unbc.ca](mailto:michalos@unbc.ca)



UNBC Ethics Committee

April 20, 2001

Bonnie Jean A. Foulds  
3290 McGill Cres  
Prince George, BC  
V2N 4E2

Proposal: 2001.0406.37

Dear Ms. Foulds:

Thank you for submitting your proposal entitled, "Gender and Relative Age Differences in Math Fluency Using Curriculum-Based Measurement."

Your proposal has been approved and you may proceed with your research.

If you have any questions, please feel free to contact me.

Sincerely,

A handwritten signature in black ink, appearing to read 'Alex', is written above the typed name.

Alex Michalos  
Chair, UNBC Ethics Review Committee

November 23, 2000 7:11:30 AM

Message

From: norm\_monroe@fc.schdist57.bc.ca

Subject: RE: RE: thesis topic

To: BJ Foulds

Hi Again BJ,

No problem with you proceeding with this research....looks good. I chatted briefly with Carl and we both agree. To get a list of those who received inservice you will probably have to call Carl or Martha Ottessen.

Best wishes....

Norm

-----Original Message-----

From: BJ\_Foulds@fc.schdist57.bc.ca

[SMTP:BJ\_Foulds@fc.schdist57.bc.ca]

Sent: Wednesday, November 22, 2000 2:54 PM

To: Norm\_Monroe@fc.schdist57.bc.ca

Subject: Re: RE: thesis topic

norm\_monroe@fc.schdist57.bc.ca writes:

>>Hi Bonnie-Jean,

>>Yes, I received your letter. My apologies for not getting back to you. Does your proposal simply involve access to the cbm data or will it involve our students?

Norm I do not need to work with/involve any students. Basically, my request is permission to access the cbm data. My project should be similar to the one done by Lynne Hedekar for the Reading and Writing CBM, with a few changes. Part of the thesis will be the comparison between the math gender and relative-age differences and those Lynne found in the Reading and Writing. What I need is

1. To have access to the SPSS program, the District bought a copy for Gail Walraven to use. It would help if I could also borrow it while doing my thesis.

2. The file with all the data from the Norming project. I will probably eliminate some students from the data set for my study, such as those that were transient during the

project. I will need info from the technical report done for the norming project also.

3. I will be doing an inter-rater reliability study. For that I need to use some of the test probes. Carl already gave me permission to use these. Will also need to know who received inservice for scoring the probes so they can be asked to participate in the inter-rater reliability part of the study. When the plans are formulated a letter from you, Martha or Carl supporting this might be helpful.

4. I could also use your assistance locating newspaper articles, journals, ministry documents, etc. with test result information that discusses gender differences, and relative-age especially for math.

Does this give you the info you wanted to know?

Thanks for your help. BJ

>>

>>Norm

>>

>>-----Original Message-----

>>From: BJ Foulds [SMTP:BJ\_Foulds@fc.schdist57.bc.ca]

>>Sent: Wednesday, November 22, 2000 12:29 PM

>>To: Norm\_Monroe@fc.schdist57.bc.ca

>>Subject: thesis topic

>>

>>Hi Norm

>>

>>A few weeks ago I sent a letter to you via school mail. I was requesting permission to use the CBM Math norming results for my thesis topic. The plan is to do gender differences and relative-age. Since I have not heard back I am wondering if you received the letter. If so, do you think I will be receiving a favourable response? I am ready to begin writing my

>>proposal so I am wondering if my topic will be a go.

>>

>>Thanks for your help.

>>

>>BJ Foulds

>>Heather Park

Appendix B  
Draft Technical Report of the CBM (Math)  
Norming Project

DRAFT TECHNICAL REPORT  
OF THE  
CURRICULUM BASED MEASUREMENT (MATH)  
NORMING PROJECT

Submitted by  
Gail B. Walraven, School District 57  
and  
Peter D. MacMillan, Ph.D.  
Education Programme  
University of Northern British Columbia

August 22, 2000

### Data Collection

Data for this project were collected from all 52 elementary schools in School District #57, Prince George. The sample for this study comprised approximately 20% of the students from each grade level (1-7). Personnel from each school received training on how to select students, how to administer and score the probes, and how to record and submit the data.

Schools used the Student Selection/Probe Sequence table (included in the handbook) to determine how many students to include. The data indicates that schools accurately selected the correct number of students. The number of students by grade for each norming period is shown in Table 1.

Table 1

#### Number of Students by Grade

Grade	October	Percent	January	Percent	April	Percent
1					282	14.55
2*	278	16.66	278	16.76	275	14.19
3	280	16.79	277	16.69	279	14.40
4	278	16.66	276	16.64	274	14.14
5	281	16.85	276	16.64	277	14.29
6	276	16.55	277	16.69	276	14.24
7	275	16.49	275	16.58	275	14.19
Total	1668	100.00	1659	100.00	1938	100.00

\*One Grade 2 student was removed from the data file. His scores were problematic; however, his school had not returned the hard copies of the probes, so verification was impossible.

The instructions in the handbook indicated which probes were to be administered at each school. The data in Table 2 suggests that the group of schools that administered Probe 3 in October, Probe 4 in January and Probe 5 in April had a slightly larger group than was expected.

Table 2

#### Number of Students by Probe

Probe	October	Percent	January	Percent	April	Percent
1	277	16.61	273	16.46	310	16.00
2	290	17.38	250	15.07	319	16.46
3	300	17.98	286	17.24	326	16.82
4	258	15.47	327	19.71	336	17.34
5	268	16.07	257	15.49	346	17.85
6	275	16.49	266	16.03	301	15.53
Total	1668	100.00	1659	100.00	1938	100.00

Probes were administered to the selected students in October, 1999, January, 2000 and April, 2000. The data collected were recorded at each school in a FileMaker Pro database file that was then forwarded electronically to Bonnie Chappell at the school board office. In May, all hard copies of the administered probes and summary data collection sheets were also forwarded to the school board office. (Some have not yet been returned)

The individual files from each school were compiled into one large file and forwarded electronically to Gail Walraven. The data were screened and cleaned before being transferred into the SPSS computer program for data analysis. Several copies of both the FileMaker Pro file and the SPSS file have been made and are in different locations to prevent loss of or damage to the data.

#### Problems in the Analysis

One school (School X) forgot to administer the April probes. They were not administered until June and this held up the final compilation of the data file. A decision had to be made as to whether data from School X would be included in the creation of the norms. The means at each grade level across all probes administered in April were compared for all schools and for all schools except School X. The results are recorded in Table 3.

Table 3

#### Comparison of April Means

<b>Grade</b>	<b>All</b>	<b>All But School X</b>
1	15.12	14.91
2	31.87	31.53
3	36.95	36.50
4	36.66	36.02
5	38.68	38.85
6	58.27	58.25
7	58.68	57.71

Based on the comparison of the means in Table 3, the decision was made to include the data from School X as there was not a significant difference in the means.

#### Demographic Analyses

The total sample consisted of 2038 students from 52 schools. There are 2039 records in the data set as one student is listed at different schools in different norming periods. Schools with students in the French Immersion or Montessori programs selected a random number of students from within these programs and submitted these data for these identified students in addition to data for students enrolled in their regular program. Grade 1 students were tested only in April. Most of the students in the sample were tested in all three norming periods. Some students were present for only one or two of the norming periods. Table 4 depicts this information.



Table 4

Students Present at Norming Periods

October	January	April	Total
√	√	√	1557
		√	317
√			50
	√	√	49
√	√		48
√		√	13
	√		5
			2039

All records submitted had complete data for gender. The data in Table 5 shows that slightly more males than females are included in the sample. This ratio of male and female students is almost the same as the ratio for the reading and writing norming study conducted by School District 57 in 1995-1996.

Table 5

Number of Students by Gender

Gender	Number	Percent
Female	997	48.9
Male	1042	51.1
Total	2039	100.0

# TECHNICAL REPORT – NORMING RESULTS

## CBM (MATH) PERCENTILE SCORES

### Smoothing

A manual smoothing process was used to create both the norming tables and the charts. Original, unsmoothed data are presented in the first set of percentile tables. This is the raw data. Data presented in both the norming tables and the charts have been smoothed. These are to be used to match a student's raw score to a percentile rank. Growth is indicated at all grade levels between norming periods. Growth is generally greater between fall and winter than between winter and spring. Also the amount of growth between norming periods is greater for the younger grades. The greatest amount of smoothing was required in Grades 6 and 7 where winter scores were higher than spring.

### GRADE ONE PERCENTILE SCORES – RAW DATA

<i>GRADE ONE Correct Digits Scored</i>				
	Fall	Winter	Spring	
Percentile	CD	CD	CD	Description
99			56	
95			35	
90			28	Well Above Average
85			24	
80			20	
75			19	Above Average
70			18	
65			17	
60			15	
55			14	
50			13	Average
45			12	
40			11	
35			10	
30			9	
25			8	Below Average
20			7	
15			6	
10			5	Well Below Average
5			4	
1			1	
n =			282	Number in Sample

N.B. Grade One students were tested only once, during the Spring norming period.

GRADE TWO PERCENTILE SCORES – RAW DATA

<i>GRADE Two Correct Digits Scored</i>				
	Fall	Winter	Spring	
Percentile	CD	CD	CD	Description
99	50	69	74	
95	36	50	55	
90	27	43	51	Well Above Average
85	23	40	48	
80	22	37	46	
75	19	33	44	Above Average
70	17	31	41	
65	15	29	38	
60	14	27	35	
55	12	25	34	
50	11	23	31	Average
45	10	21	29	
40	9	19	27	
35	8	18	26	
30	7	17	23	
25	6	15	20	Below Average
20	6	13	17	
15	5	11	15	
10	4	9	12	
5	2	6	8	
1	0	2	3	
n =	278	278	275	Number in Sample

GRADE THREE PERCENTILE SCORES – RAW DATA

<b>GRADE THREE Correct Digits Scored</b>				
	<b>Fall</b>	<b>Winter</b>	<b>Spring</b>	
<b>Percentile</b>	<b>CD</b>	<b>CD</b>	<b>CD</b>	<b>Description</b>
99	53	61	71	
95	45	54	61	
90	40	51	56	<b>Well Above Average</b>
85	35	46	53	
80	30	43	50	
75	28	40	47	<b>Above Average</b>
70	26	38	45	
65	24	36	43	
60	22	34	41	
55	20	32	39	
50	19	28	37	<b>Average</b>
45	18	27	35	
40	17	25	32	
35	16	24	30	
30	14	22	28	
25	13	21	26	<b>Below Average</b>
20	11	19	24	
15	9	17	22	
10	6	16	19	<b>Well Below Average</b>
5	3	11	15	
1	0	4	4	
n =	280	277	279	Number in Sample

GRADE FOUR PERCENTILE SCORES – RAW DATA

<b>GRADE FOUR Correct Digits Scored</b>				
	<b>Fall</b>	<b>Winter</b>	<b>Spring</b>	
<b>Percentile</b>	<b>CD</b>	<b>CD</b>	<b>CD</b>	<b>Description</b>
99	51	75	84	
95	39	60	69	
90	33	54	59	<b>Well Above Average</b>
85	29	48	55	
80	27	46	51	
75	26	41	47	<b>Above Average</b>
70	25	39	44	
65	22	36	42	
60	21	34	38	
55	20	32	36	
50	19	29	33	<b>Average</b>
45	17	28	32	
40	16	26	31	
35	14	24	29	
30	13	23	27	
25	12	21	25	<b>Below Average</b>
20	10	18	22	
15	8	16	18	
10	6	12	14	<b>Well Below Average</b>
5	5	9	10	
1	3	2	3	
n =	278	276	274	Number in Sample

GRADE FIVE PERCENTILE SCORES – RAW DATA

<i><b>GRADE FIVE Correct Digits Scored</b></i>				
	<b>Fall</b>	<b>Winter</b>	<b>Spring</b>	
<b>Percentile</b>	<b>CD</b>	<b>CD</b>	<b>CD</b>	<b>Description</b>
99	60	88	91	
95	45	63	72	
90	40	55	63	<b>Well Above Average</b>
85	34	50	59	
80	31	48	54	
75	28	46	50	<b>Above Average</b>
70	26	43	47	
65	24	40	45	
60	23	37	43	
55	21	35	39	
50	20	32	37	<b>Average</b>
45	19	30	35	
40	17	29	33	
35	16	26	30	
30	15	25	27	
25	14	22	24	<b>Below Average</b>
20	13	20	21	
15	12	18	19	
10	9	13	15	<b>Well Below Average</b>
5	8	11	12	
1	3	6	5	
n =	281	276	277	Number in Sample

GRADE SIX PERCENTILE SCORES – RAW DATA

<i>GRADE SIX Correct Digits Scored</i>				
Percentile	Fall CD	Winter CD	Spring CD	Description
99	97	116	123	
95	84	96	99	
90	71	84	91	Well Above Average
85	62	77	86	
80	58	72	80	
75	54	68	74	Above Average
70	50	65	69	
65	48	63	65	
60	45	60	61	
55	43	58	58	
50	40	54	56	Average
45	38	50	53	
40	36	47	50	
35	34	45	46	
30	32	42	44	
25	29	39	40	Below Average
20	27	36	37	
15	25	34	34	
10	23	29	29	Well Below Average
5	17	24	23	
1	9	19	15	
n =	276	277	276	Number in Sample



GRADE SEVEN PERCENTILE SCORES – RAW DATA

<i>GRADE SEVEN Correct Digits Scored</i>				
	Fall	Winter	Spring	
<b>Percentile</b>	<b>CD</b>	<b>CD</b>	<b>CD</b>	<b>Description</b>
99	123	130	134	
95	86	99	102	
<b>90</b>	<b>78</b>	<b>88</b>	<b>89</b>	<b>Well Above Average</b>
85	74	81	82	
80	69	75	78	
<b>75</b>	<b>63</b>	<b>70</b>	<b>72</b>	<b>Above Average</b>
70	59	66	70	
65	54	64	67	
60	51	60	65	
55	48	57	60	
<b>50</b>	<b>45</b>	<b>53</b>	<b>57</b>	<b>Average</b>
45	42	50	53	
40	38	47	51	
35	36	44	48	
30	32	41	46	
<b>25</b>	<b>30</b>	<b>37</b>	<b>41</b>	<b>Below Average</b>
20	27	33	38	
15	24	29	32	
<b>10</b>	<b>20</b>	<b>25</b>	<b>26</b>	<b>Well Below Average</b>
5	14	20	19	
1	6	9	9	
n =	275	275	275	Number in Sample



## APPENDIX A:

## SUMMARY RESULTS

Descriptive StatisticsGrade One Results

	<b>Mean</b>	<b>S.D.</b>	<b>Min</b>	<b>Max</b>	<b>Skew</b>	<b>Kurtosis</b>
<b>Fall</b>	-	-	-	-	-	-
<b>Winter</b>	-	-	-	-	-	-
<b>Spring</b>	15.12	10.28	0	63	1.69	3.93

Grade Two Results

	<b>Mean</b>	<b>S.D.</b>	<b>Min</b>	<b>Max</b>	<b>Skew</b>	<b>Kurtosis</b>
<b>Fall</b>	13.90	10.29	0	59	1.41	2.42
<b>Winter</b>	25.14	13.73	1	77	0.76	0.68
<b>Spring</b>	31.87	15.18	1	80	0.27	-0.23

Grade Three Results

	<b>Mean</b>	<b>S.D.</b>	<b>Min</b>	<b>Max</b>	<b>Skew</b>	<b>Kurtosis</b>
<b>Fall</b>	21.20	12.10	0	56	0.62	-0.05
<b>Winter</b>	30.86	13.33	0	66	0.30	-0.62
<b>Spring</b>	36.95	14.41	0	81	0.12	-0.35

Grade Four Results

	<b>Mean</b>	<b>S.D.</b>	<b>Min</b>	<b>Max</b>	<b>Skew</b>	<b>Kurtosis</b>
<b>Fall</b>	19.38	10.49	1	71	0.87	1.67
<b>Winter</b>	31.74	15.89	1	87	0.61	0.20
<b>Spring</b>	36.53	17.95	0	134	0.89	2.49

Grade Five Results

	<b>Mean</b>	<b>S.D.</b>	<b>Min</b>	<b>Max</b>	<b>Skew</b>	<b>Kurtosis</b>
<b>Fall</b>	22.23	11.54	1	82	1.16	2.42
<b>Winter</b>	34.45	16.48	5	98	0.68	.80
<b>Spring</b>	38.70	18.76	0	102	0.53	0.13

Grade Six Results

	<b>Mean</b>	<b>S.D.</b>	<b>Min</b>	<b>Max</b>	<b>Skew</b>	<b>Kurtosis</b>
<b>Fall</b>	43.41	19.03	5	102	0.75	0.32
<b>Winter</b>	55.47	21.33	13	124	0.55	0.06
<b>Spring</b>	58.27	23.93	7	129	0.45	-0.24

Grade Seven Results

	<b>Mean</b>	<b>S.D.</b>	<b>Min</b>	<b>Max</b>	<b>Skew</b>	<b>Kurtosis</b>
<b>Fall</b>	47.54	23.15	4	132	0.61	0.38
<b>Winter</b>	55.36	24.96	8	148	0.58	0.47
<b>Spring</b>	58.68	24.97	5	159	0.53	0.75

### Technical Adequacy

Previous research done on CBM measures of reading fluency and written expression has reported that CBM measures have demonstrated stability over time and across testers. School District 57's 1995-1996 norming project for reading and written expression endorsed this.

However, very little research has yet been done on CBM in math. The results of the Pearson Correlation in the following table for correct digits scored compared between norming periods are stable. They indicate stability over time (6 months), and equivalence of the probes. As stability is present across groups, it can be assumed that results would be stable for an individual student.

This is evidence that the probes are indeed measuring mathematics computational skills.

### Correlations across Norming Periods

Pearson Correlation for <u>Correct Digits Scored</u> Scores between Norming Periods			
Grade	r Oct-Jan	r Jan-Apr	r Oct-Apr
1	—	—	—
2	.71	.73	.63
3	.71	.74	.65
4	.68	.74	.65
5	.53	.63	.59
6	.58	.69	.45
7	.68	.73	.60

### Analysis of Probe Difficulty

The analysis of probe difficulty is of prime importance in this project. If the probes are not of similar difficulty, they cannot be used to assess student progress. If a student were to be tested using an easier probe after a more difficult one, the measure of the progress would be exaggerated. Conversely, an underestimation of progress would occur if a more difficult probe were used after an easier one. Four techniques were used to analyze the probes for difficulty.

The probe difficulties for each grade level were examined using a one way ANOVA. This was followed by the Scheffé post hoc comparison using  $\alpha < .01$ , if the ANOVA omnibus test results indicated significant differences. This procedure was selected as it will provide a relatively low number of false positives. It is not as likely to claim probes are of different difficulties when they are in fact of equal difficulty. Where significant differences were found, the probe order was compared across norming periods. Probes are considered to be candidates for being different only when the same probe is consistently found to differ in a consistent manner from the other probes at that grade level. Finally, probes were subjected to a very conservative test for evidence of probe differences. Box plots of the probes for each grade level were examined for lack of overlap.

In the following tables, the notation “ns” is used to indicate no statistically significant differences, while the notation “sig” is used to indicate significant differences have been found using the Scheffé post hoc comparison. A short interpretation is provided after each table.

## Probe Difficulty

Table 1  
Math Probe Differences

Grade 1 CD	Probe OCT	Probe JAN	Probe APR	Mean OCT	Mean JAN	Mean APR
			ns			ns
			2			12.43
			3			13.90
			5			14.41
			6			15.47
			4			15.94
			1			18.78

No probes were judged significantly different at the Grade 1 level.

Table 2  
Math Probe Differences Across Norming Periods

Grade 2 CD	Probe OCT	Probe JAN	Probe APR	Mean OCT	Mean JAN	Mean APR
	sig	sig	ns	sig	sig	ns
	1	3	6	11.26	18.94	28.42
	4	2	3	12.00	22.74	29.00
	2	5	4	12.73	24.23	30.38
	3	1	2	13.22	24.80	31.73
	6	4	5	13.76	29.58	33.14
	5	6	1	20.77	29.77	38.77

No probes were judged significantly different at the Grade 2 level. Probe 5 appears to be significantly easier than the others in October; however this does not hold over the other two norming periods.

Table 3  
Math Probe Differences Across Norming Periods

Grade 3 CD	Probe OCT	Probe JAN	Probe APR	Mean OCT	Mean JAN	Mean APR
	sig	sig	sig	sig	sig	sig
	4	5	6	14.51	26.95	31.05
	1	2	2	15.48	27.09	34.42
	6	1	4	19.78	27.24	36.73
	5	4	5	23.09	28.51	38.38
	2	6	3	26.18	36.36	39.51
	3	3	1	27.14	39.00	41.11

No probes were judged significantly different at the Grade 3 level.

Table 4  
Math Probe Differences Across Norming Periods

Grade 4 CD	Probe OCT	Probe JAN	Probe APR	Mean OCT	Mean JAN	Mean APR
	ns	sig	ns	ns	sig	ns
	2	5	3	18.33	25.86	31.42
	6	3	6	18.46	27.94	34.67
	5	2	5	18.51	30.12	35.72
	3	6	2	19.38	34.07	36.85
	4	4	4	20.32	34.38	37.72
	1	1	1	21.40	37.20	42.89

At the Grade 4 level, it appears that Probe 1 and 4 are consistently significantly easier.

Table 5  
Math Probe Differences Across Norming Periods

Grade 5 CD	Probe OCT	Probe JAN	Probe APR	Mean OCT	Mean JAN	Mean APR
	sig	sig	sig	sig	sig	sig
	5	5	5	15.62	27.58	31.16
	1	3	3	20.69	31.50	37.44
	3	4	6	21.59	34.60	37.65
	2	2	4	23.53	36.07	39.29
	6	6	1	25.06	36.73	41.20
	4	1	2	26.86	39.92	45.65

At the Grade 5 level, Probe 5 appears significantly easier.

Table 6  
Math Probe Differences Across Norming Periods

Grade 6 CD	Probe OCT	Probe JAN	Probe APR	Mean OCT	Mean JAN	Mean APR
	sig	sig	sig	sig	sig	sig
	5	5	5	35.84	42.74	48.51
	4	6	6	37.33	50.96	49.09
	6	4	4	40.13	52.98	52.27
	1	3	3	45.43	58.02	60.22
	3	2	1	49.82	61.68	65.57
	2	1	2	50.50	66.82	74.57

At the Grade 6 level, it appears that Probes 4, 5 and 6 are consistently more difficult than Probes 1, 2 and 3.

Table 7

Math Probe Differences Across Norming Periods

Grade 7 CD	Probe OCT	Probe JAN	Probe APR	Mean OCT	Mean JAN	Mean APR
	sig	sig	ns	sig	sig	ns
	4	2	2	39.14	44.44	52.95
	2	5	6	40.46	48.37	54.23
	1	3	1	45.37	56.14	57.29
	3	1	3	48.80	57.64	57.33
	6	6	4	53.95	60.30	63.43
	5	4	5	57.91	62.60	65.41

No probes were judged significantly different at the Grade 7 level.

Summary of Probe Difficulty

As indicated in the interpretations of the tables, the strongest evidence of probe difficulty was at the Grade 4, 5 and 6 levels. However, in examining the box plots for these grade levels, no lack of overlap was found at the Grade 4 level. A lack of overlap was present for only one norming period for both Grades 5 and 6. Probes may be judged equal for Grades 1-3 and Grade 7. The committee may wish to examine the identified probes at the Grades 4-6 levels.

**Appendix C**  
**Draft Math Norms Tables for**  
**Curriculum Based Measurement Calculation**

**(Note: The administration and scoring rules provided with the norms tables are available in Appendix F with the CBM Math Training Norming Project hand-outs)**



**DRAFT**

**School District No. 57**

**Norms Tables**

**for**

**Curriculum Based Measurement  
Math Calculation**

**September 18, 2000**

### **Development of Curriculum Based Measurement (CBM) Norms in School District 57 (Prince George)**

In the spring of 1995, a joint School District 57 - University of Northern British Columbia (UNBC) project to develop local CBM norms at the elementary level for reading and written expression was established. Testing procedures, materials and teacher inservice plans were developed and the norming project was implemented during the 1995-96 school year. CBM reading and written expression probes were administered by S.D. 57 teachers to randomly selected students. A UNBC professor and graduate student processed the data, developed norms tables and technical reports. The Guidebook for the Use of Curriculum Based Measurement in School District # 57 was presented to schools at an inservice in November, 1996.

Within a few years the need for a similar standardized, norm-referenced assessment tool in mathematics at the elementary level was identified. During the spring and summer of 1995, a joint School District 57 - University of Northern British Columbia (UNBC) project to develop local CBM norms for elementary math calculations was developed. A training inservice was held in September, 2000. Three times during the 1999 - 2000 school year three CBM math calculation probes were administered by S.D. 57 teachers to randomly selected students. The data were processed, norms tables and technical reports were created at UNBC.

The math calculation norms tables are being made available for use in schools starting September, 2000. The norming project probes and the instructions for administration and scoring should be used in conjunction with the norms tables.

A guidebook for the use of Curriculum Based Measurement - Math Calculations is under development and will be presented to schools later this fall.

September 18, 2000

GRADE ONE NORMS

Percentile	CD	CD	CD	Description
95			35	
85			24	
75			19	Above Average
65			17	
55			14	
45			12	
35			10	
25			8	Below Average
15			6	
5			4	

N.B. Grade One students were tested only once, during the Spring norming period.

GRADE TWO NORMS

Percentile	CD	CD	CD	Description
95	36	50	60	
85	23	40	48	
75	19	33	44	Above Average
65	15	29	38	
55	12	25	34	
45	10	21	29	
35	8	18	26	
25	6	15	20	Below Average
15	4	11	15	
5	2	6	8	

GRADE THREE NORMS

Percentile	CD	CD	CD	Description
95	45	54	61	
85	35	46	53	
75	28	40	47	Above Average
65	24	36	43	
55	20	32	39	
45	18	27	35	
35	16	24	30	
25	13	21	26	Below Average
15	9	17	22	
5	3	10	15	

GRADE FOUR NORMS

Percentile	CD	CD	CD	Description
95	39	60	69	
85	29	48	55	
75	26	41	47	Above Average
65	22	36	41	
55	20	32	36	
45	17	28	32	
35	14	24	29	
25	12	21	25	Below Average
15	8	15	18	
5	5	8	11	

GRADE FIVE NORMS

Percentile	CD	CD	CD	Description
95	45	63	72	
85	34	50	59	
75	28	46	50	Above Average
65	24	40	45	
55	21	35	39	
45	19	30	35	
35	16	26	30	
25	14	22	25	Below Average
15	12	16	19	
5	8	10	12	

GRADE SIX NORMS

Percentile	CD	CD	CD	Description
95	84	96	101	
85	62	77	86	
75	54	68	74	Above Average
65	48	63	65	
55	43	57	59	
45	38	50	53	
35	34	44	47	
25	29	38	41	Below Average
15	25	30	34	
5	17	22	24	



GRADE SEVEN NORMS

Percentile	CD	CD	CD	Description
95	86	99	110	
85	74	80	84	
75	63	69	72	Above Average
65	54	64	67	
55	48	57	60	
45	42	50	53	
35	36	44	48	
25	30	37	41	Below Average
15	24	29	32	
5	14	17	20	

## Appendix D

### Sample CBM Math Probe and Answer Key

Grade 7, probe 1

PEN \_\_\_\_\_ Name \_\_\_\_\_ CD \_\_\_\_\_

$\begin{array}{r} 15637 \\ - 9859 \\ \hline \end{array}$	$(+9) - (-4) = \underline{\hspace{2cm}}$	$\begin{array}{r} 19 \\ \times 98 \\ \hline \end{array}$	$\begin{array}{r} 36 \\ 47 \\ + 12 \\ \hline \end{array}$	$72 \overline{)1651}$
$14.2 + 24.7 = \underline{\hspace{2cm}}$	$.6 \overline{)7.32}$	$26 \overline{)403}$	$(-7) - (-2) = \underline{\hspace{2cm}}$	$10.4 + 9.12 = \underline{\hspace{2cm}}$
$4 \overline{)3.48}$	$57 \div \underline{\hspace{2cm}} = 3$	$(+7) \times (-3) = \underline{\hspace{2cm}}$	$\frac{24}{6} =$	$(+5) \div (-1) = \underline{\hspace{2cm}}$

Grade 7, probe 1

$\begin{array}{r} 587 \\ \times 37 \\ \hline \end{array}$	$4.2 - 1.58 = \underline{\hspace{2cm}}$	$\begin{array}{r} 8637 \\ - 2918 \\ \hline \end{array}$	$12\% \times 50 = \underline{\hspace{2cm}}$	$\begin{array}{r} 179.4 \\ + 25.6 \\ \hline \end{array}$
$\begin{array}{r} 8437 \\ + 5976 \\ \hline \end{array}$	$(+8) + (-2) = \underline{\hspace{2cm}}$	$\underline{\hspace{2cm}} \times 70 = 4900$	$(+3) + (+4) = \underline{\hspace{2cm}}$	$\begin{array}{r} \phantom{0} \\ 32 \overline{)608} \\ \hline \end{array}$
$\begin{array}{r} 2303 \\ - 1805 \\ \hline \end{array}$	$.174 = \underline{\hspace{2cm}} \%$	$14 \times \underline{\hspace{2cm}} = 700$	$3600 \div 60 = \underline{\hspace{2cm}}$	$20\% \text{ of } 70 = \underline{\hspace{2cm}}$

Grade 7, probe 1

PEN \_\_\_\_\_ Name \_\_\_\_\_ CD \_\_\_\_\_

151

$\begin{array}{r} 15637 \\ - 9859 \\ \hline 5778 \end{array}$	$(+9) - (-4) = \overset{+}{13}$ or 13	$\begin{array}{r} 19 \\ \times 98 \\ \hline 152 \\ 1710 \\ \hline 1862 \end{array}$	$\begin{array}{r} 36 \\ 47 \\ + 12 \\ \hline 95 \end{array}$	$\begin{array}{r} 22 \\ 72 \overline{)1651} \\ \underline{144} \\ 211 \\ \underline{144} \\ 67 \end{array}$	
(4)	(3)	(11)	(2)	(13)	33
$14.2 + 24.7 = \underline{38.9}$	$\begin{array}{r} 12.2 \\ 6 \overline{)7.32} \\ \underline{6} \\ 13 \\ \underline{12} \\ 12 \\ \underline{12} \\ 0 \end{array}$	$\begin{array}{r} 15 \\ 26 \overline{)403} \\ \underline{26} \\ 143 \\ \underline{130} \\ 13 \end{array}$	$(-7) - (-2) = \underline{-5}$	$10.4 + 9.12 = \underline{19.52}$	
(4)	(14)	(12)	(2)	(5)	37
$\begin{array}{r} .87 \\ 4 \overline{)3.48} \\ \underline{32} \\ 28 \\ \underline{28} \\ 0 \end{array}$	$57 \div \underline{19} = 3$	$(+7) \times (-3) = \underline{-21}$	$\frac{24}{6} = 4$	$(+5) \div (-1) = \underline{-5}$	
(10)	(2)	(3)	(1)	(2)	18

Grade 7, probe 1

$\begin{array}{r} 587 \\ \times 37 \\ \hline 4109 \\ 17610 \\ \hline 21719 \end{array}$	$4.2 - 1.58 = \underline{2.62}$	$\begin{array}{r} 8637 \\ - 2918 \\ \hline 5719 \end{array}$	$12\% \times 50 = \underline{6}$	$\begin{array}{r} 179.4 \\ + 25.6 \\ \hline 205.0 \\ \text{or} \\ 205 \end{array}$	
(14)	(4)	(4)	(1)	(5)	<u>28</u>
$\begin{array}{r} 8437 \\ + 5976 \\ \hline 14413 \end{array}$	$(+8) + (-2) = \underline{+6}$	$\underline{70} \times 70 = 4900$	$(+3) + (+4) = \underline{7}$ or 7	$\begin{array}{r} 19 \\ 32 \overline{)608} \\ \underline{32} \\ 288 \\ \underline{288} \\ 0 \end{array}$	
(5)	(2)	(2)	(2)	(11)	<u>22</u>
$\begin{array}{r} 2303 \\ - 1805 \\ \hline 498 \end{array}$	$.174 = \underline{17.4\%}$	$14 \times \underline{50} = 700$	$3600 \div 60 = \underline{60}$	$20\% \text{ of } 70 = \underline{14}$	
(3)	(4)	(2)	(2)	(2)	

## Appendix E

The CBM Math (Calculation) Norming Training Project, 1999-2000 Hand-outs

**School District #57  
Curriculum Based Measurement Norming Project**

**TIMELINE**

**SEPTEMBER 22 ,1999**

Inservice for Administrators and Teachers

**Norming Period #1**

**SEPTEMBER 27, 1999**

Random Selection of students for project, Grade 2 - 7

**OCTOBER 4 -15, 1999**

Do calculation probes, Grades 2 - 7

**OCTOBER 22, 1999**

Deadline for submitting data via 57Online.  
Send student probes organized by grade level to  
Sharon Priseman at Central Administration Building.

**Norming Period #2**

**JANUARY 17 - 28, 2000**

Do calculation probes, Grades 2 - 7

**FEBRUARY 4, 2000**

Deadline for submitting data via 57Online.  
Send student probes organized by grade level to  
Sharon Priseman at Central Administration Building.

**Norming Period #3**

**APRIL 13, 2000**

Random Selection of students for project, Grade 1

**APRIL 17 - 28, 2000**

Do calculation probes, Grades ~~2~~ 1 - 7

**MAY 5, 2000**

Deadline for submitting data via 57Online.  
Send student probes *and data recording forms*  
organized by grade level to Sharon Priseman at Central  
Administration Building.



### Norming Procedures

#### 1. Arrange Students by Grade Level

On September 27, 1999, generate an alphabetical list of students in each grade (2 - 7). On the list, indicate if the student is First Nations, is enrolled in French Immersion or Montessori.

**Note:** Exclude Grade 1 students during the October and January norming periods. Perform the above steps with Grade 1 students on April 13, 2000.

#### 2. Apply Exclusion Criteria

Exclude students from the lists who fit under the following categories:

- a) Level 1 & 2 ESL students
- b) Students with intellectual disabilities
- c) Other "hard labeled" students (hearing impaired, visually impaired, autistic, multiply disabled)

#### 3. Select Students at Random

For each list of names use the Random Selection of Students Form to determine which students from the list correspond with the random numbers generated for that particular grade level at your school.

1. If the random number is greater than the number of names on a list:
  - a) Count all the names on the list
  - b) Go to the beginning of the list and continue counting until the number in question is reached - the student name which corresponds with this number is the student selected.
2. If the random number corresponds with a student already selected for the norming sample:
  - a) Roll a die
  - b) If the number on the die is even, the next available student higher on the list is selected.
  - c) If the number on the die is odd, the next available student lower on the list is selected.
3. Continue selecting students until you have selected the number per grade assigned to your school. Refer to Student Selection / Probe Sequence Chart.

#### 4. Complete Student Information on CBM Data Recording Form

Record the information requested for each student selected on the CBM Math Probe (Calculation) Data Recording Form. Use a separate form for each grade level. After recording the information, double check for accuracy.

### **5. Administer Calculation Probes**

1. During the first norming period, at each grade level, administer the probe number listed for your school. Refer to Student Selection/Probe Sequence Chart. During the next norming periods, administer the next probes in numerical sequence. (eg. October = probe 3, January = probe 4, April = probe 5)

2. Norming Periods

Refer to the Timeline for information

3. *Grade 1 students*

Grade 1 students are excluded from norming activities during the October and January norming periods. *They are included during the April norming activities.*

4. Use the administration procedures included in this manual.

### **6. Score Calculation Probes**

Use the scoring procedures included in this manual. These scoring procedures are based on the CBM Training Manual edited by Mark Shinn, Nancy Knutson, and David Tilly.

### **7. Record Scores on CBM Data Recording Form**

1. After writing this information down on the CBM Math Probe (Calculation) Data Recording Form, double check it for accuracy.
2. Make and keep a photocopy of the completed recording form.

### **8. Transfer the information on the CBM Data Recording Form to (.... 57 Online). Organize student probes by grade level and send to Sharon Priseman at the Central Administration Building.**

1. Contact *Bonnie Chappell*, at the Central Administration Building, if you have questions or difficulties with the transfer process.
2. At the end of the April norming period, send all data recording forms used in this norming project to *Sharon Priseman* at the Central Administration Building. Organize the forms by grade level.

### Student Selection/Probe Sequence

ELEMENTARY SCHOOLS	(per grade)	Probe #
Morlee	8	1
Austin Road	9	1
Immerision	2	1
Varway	6	1
Springwood	7	1
Van Buren	5	1
Meadow	4	1
Seymour	3	1
Montessori	gr.1-4, gr.2=2	1
Red Rock	1	1
Dunster	1	1
Westwood	10	2
Carney Hill	8	2
Beverly	7	2
Lakewood	6	2
Glenview	5	2
Fl. George South	3	2
Edgewood	2	2
Dome Creek	0	2
Pineview	6	2
Hixon	1	2
Highglen	3	3
Montessori	5	3
Mackenzie	7	3
Highland	6	3
College Heights	5	3
Immerision	4	3
Southridge	9	3
Wildwood	5	3
K.G.V. English	3	3
Haldi Road	2	3
McLeod Lake	1	3
Bear Lake	1	4
Blackburn	11	4

ELEMENTARY SCHOOLS	(per grade)	Probe #
Quinson	6	4
Ron Brent	5	4
North Nechako	6	4
Mountain View	5	4
Harwin	4	4
Central Fort George	3	4
Giscome	2	4
Valemount	6	5
Hart Highlands	9	5
Heritage	8	5
Peden Hill	7	5
Gladstone	6	5
McBride Centennial	5	5
Buckhorn	4	5
Malaspina	8	6
Foothills	7	6
Hart Highway	7	6
Spruceland	6	6
Immerision	4	6
Nukko Lake	5	6
Pinewood	4	6
Shady Valley	3	6
Salmon Valley	2	6

### Random Selection Of Students

	Students to be used in Norming Sample
<b>Grade 1</b>	10, 11, 15, 18, 19, 23, 27, 28, 30, 34, 41
<b>Grade 2</b>	4, 5, 7, 8, 16, 18, 20, 22, 23, 35, 47
<b>Grade 3</b>	2, 3, 16, 22, 26, 27, 33, 37, 41, 45, 51
<b>Grade 4</b>	4, 11, 12, 13, 25, 26, 30, 39, 40, 43, 50
<b>Grade 5</b>	4, 5, 7, 18, 20, 21   23, 28, 34, 36, 42
<b>Grade 6</b>	7, 8, 12, 13, 15, 19,   23, 28, 30, 39, 44
<b>Grade 7</b>	5, 8, 9, 13, 21, 26   37, 39, 41, 45, 53

### **HOW TO INFORM PARENTS OF THE NORMING PROJECT**

#### **Suggested Newsletter Insert:**

School District #57 will be collecting math calculation samples from elementary students three times during this school year, once in October, once in January, and once in April. The data collected will be used to create statistical tables showing the range of student performance at each grade level. Participating students are chosen at random and will remain anonymous. If you have questions, please contact your school principal.

### **WHAT TO TELL THE STUDENTS ABOUT THE PROJECT**

Tell students that School District #57 is collecting math samples from 300 children in each grade three times this year, once in October, once in January, and once in April. These samples will provide information about how children in this district add, subtract, multiply, and divide. All children are chosen at random (explain this if necessary) and will remain anonymous (explain this if necessary).

### **WHAT TO DO IF...**

#### **A TARGET STUDENT IS AWAY FOR ENTIRE TWO-WEEK NORMING PERIOD:**

Record no score for the student. Include the student in the next norming period.

#### **A TARGET STUDENT MOVES AWAY FROM YOUR SCHOOL:**

Generate a list of students at that grade level *who are new since September 27, 1999* to your school, and *randomly* select an alternate student. If there are no new students at that grade level, *randomly* select an alternate student from the general grade level population. Record the alternate student's information on the appropriate CBM Math Probe (Calculation) Data Recording Form.

**THE TEACHER WHO ADMINISTERED THE FIRST SET OF PROBES IS UNAVAILABLE TO DO THE NEXT SET OF PROBES:**

Inform School Services 1 month prior to norming period.

A training session for the replacement teacher will be provided.

**WHAT TO DO IF....**

**DURING A CALCULATION PROBE, A STUDENT STOPS WORKING BEFORE THE END OF THE TIME LIMIT:**

Quietly say to the student "Keep working until I tell you to stop."

**DURING A CALCULATION PROBE, A STUDENT COMPLETES ALL PROBLEMS BEFORE THE END OF THE TIME LIMIT:**

Quietly give the student the "extra" probe for that grade level. *The extra probe will be scored only if the student has answered all items on the first probe.*

**WHEN SCORING A PROBE, THE STUDENT'S WORK IS HARD TO READ :**

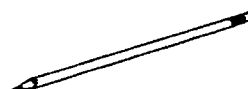
Count each digit you can read. If it is not possible to read any digits, CD score is 0.

**YOU HAVE A QUESTION THAT ISN'T ANSWERED HERE:**

Phone:

Martha Ottesen (562-8051)

### CBM Math Probe Data Recording Form in Filemaker Pro



In order to enter the data on the computer form you need to be working on the computer in your school which has Filemaker Pro installed on it, and 57 Online. Then just follow the instructions below.

#### Downloading the File from 57Online and Using Filemaker Pro

1. Logon to **57 Online**
2. Open **57 Information Centre**
3. Open **Forms and Documents**
4. Open message **CBM Math Data Collection Form**  
 To save the attachment click twice on it and then decide where on your computer you want to save it.
5. Open the program **Filemaker Pro**
6. Open the attachment **CBMMathData.fp3**  
 which you saved from 57Online somewhere on your computer.
7. Go to **FILE** menu, drag down to **SAVE COPY AS** and rename the file  
 "YourSchoolName.fp3" (eg. Haldi.fp3, or Morfee.fp3) *Quinson.fp3*



#### Setting up to Enter the Data on the Collection Sheet

1. Choose **MODE** in the top row, pull down and choose **BROWSE** (if it isn't already)
2. Choose **RECORD** button in top left below **FILE**, drag down and choose **SCHOOL**. You will see a new layout.
3. Put arrow in white area below the word school in the new layout and click. When list of school names pops up, scroll down to your school name and click on it.
4. Go back to **SCHOOL** button in top left below **FILE**, drag down and choose **RECORD** form. You will return to the Record Form Layout.

#### Entering the Data

If familiar with exporting information from Turbo school, you could export PEN, Student Names, Grade, Sex, Birthdate, and First Nations information. Then delete the students you don't need.  
 OR you can enter the data manually.

*sent*

1. Put cursor in box under **PEN** and type in Personal Education Number, then press **TAB** to move to the next column, student name and so on. Filemaker Pro saves the data as you go along so you do not need to worry about saving the data.
2. Type in Student Name, Birthdate (day/month/year), M or F for Sex, Grade, Y if First Nations, French Immersion or Montessori or leave blank.
3. Then type in the score for the October/ January/or April Probe, and the Probe Number
4. Continue until all students are entered. If you run out of room and need more records, go to **MODE** and select **NEW RECORD**

#### Sending data to Board Office

1. Open 57Online, and open a new message
2. Send to **Bonnie Chappell**
3. Subject **CBM Math Data, Your school name**
4. Go up to **FILE** and pull down to **ATTACH FILE**
5. Then go to where you saved the Filemaker Pro File and **SAVE** it
6. Send the message! You're done!



The same file can be used for the next set of records in January and April.

## School

**PEN**

**Surname, First Name**

Birthdate (m/d/y)

**Sex**

Gr.

## First Nations

## French Immers

**Montessori**

Oct CD  
score

**Oct  
Probe**

**Jan CD  
score**

## Jan Probe

**Apr CD  
score**

**April Probe**

### Probe

[illegible]



### Directions for 5 Minute Administration of Calculation Probes

#### Materials:

1. Calculation Probes
2. Scoring Template
3. Stop watch

#### Directions:

1. Provide the student with a pencil and the calculation probe with the student's name written on the top of page one. Place the probe face up on the desk in front of the student.

2. Say these specific directions to the student:

- **"The paper on your desk has several types of problems. Some are... (name types of problems included on the probe). Look at each problem carefully to decide whether to add, subtract, (multiply or divide)."**

- **"When I say 'begin' start answering the problems. Begin with the first problem and work across the page (demonstrate by pointing). Then go to the next row. Try to do every problem. If you finish one side, turn the paper over and continue working. If you finish both sides, raise your hand. You will have 5 minutes to work. Are there any questions?" (pause)**

3. Say **"Begin"** and start your stop watch.

4. Monitor students to ensure they work across the page and do not skip around or answer only specific problems. If they do, say **"Be sure to work across the page. Try to do every problem."**

**Note:** If a student completes every problem on both pages before the 5 minute time limit, give the student the "extra" probe for that grade level.

5. At the end of 5 minutes say, **"Stop. Put your pencil down."**

Administration and scoring procedures adapted from  
Administration and Scoring of Curriculum-Based Measurement, 1992,  
M. Shinn, N. Knutson, and W. Tilly III

### Scoring Rules for Math

**Correct Digits.** Credit is earned for each digit that is correct within a student response. For ease of scoring, underline all correct digits used to work out longer problems.

#### Scoring Template

$$\begin{array}{r} 8 \\ +8 \\ \hline 16 \end{array} \quad (2)$$

$$\begin{array}{r} 19 \\ -5 \\ \hline 14 \end{array} \quad (2)$$

$$\begin{array}{r} 9 \\ \times 8 \\ \hline 72 \end{array} \quad (2)$$

$$\begin{array}{r} 12 \\ 12 \overline{)144} \end{array} \quad (2)$$

#### Student Response

$$\begin{array}{r} 8 \\ +8 \\ \hline 16 \end{array} \quad (2)$$

$$\begin{array}{r} 19 \\ -5 \\ \hline 14 \end{array} \quad (2)$$

$$\begin{array}{r} 9 \\ \times 8 \\ \hline 62 \end{array} \quad (1)$$

$$\begin{array}{r} 12 \\ 12 \overline{)144} \end{array} \quad (2)$$

---

**Rule 1. Incomplete Problems.** When a student has not completed a problem, credit is earned for the correct digits written.

#### Scoring Template

$$\begin{array}{r} 32 \\ \times 15 \\ \hline 160 \\ 320 \\ \hline 480 \end{array} \quad (9)$$

#### Student Response

$$\begin{array}{r} 32 \\ \times 15 \\ \hline 160 \\ 32 \\ \hline \end{array} \quad (5)$$


---

**Rule 2. Crossed-Out Problems.** If the student has crossed-out a problem, credit is earned for the correct digits written.

Scoring Template

$$\begin{array}{r} 12 \\ \times 14 \\ \hline 48 \\ 120 \\ \hline 168 \end{array} \quad (8)$$

Student Response

$$\begin{array}{r} 12 \\ \times 14 \\ \hline 48 \\ 120 \\ \hline 168 \end{array} \quad (8)$$

**Rule 3. Carries and Borrows.** "Carries" and "borrows" are not counted as correct digits.

Scoring Template

$$\begin{array}{r} 723 \\ -564 \\ \hline 159 \end{array} \quad (3)$$

Student Response

$$\begin{array}{r} 11 \\ 6723 \\ -564 \\ \hline 159 \end{array} \quad (3)$$

**Rule 4. Alignment for Correct Answers.** If the answer is correct, the digits do not have to be correctly aligned to earn full credit.

Scoring Template

$$\begin{array}{r} 15 \\ \times 12 \\ \hline 30 \\ 150 \\ \hline 180 \end{array} \quad (8)$$

Student Response

$$\begin{array}{r} 15 \\ \times 12 \\ \hline 20 \\ 150 \\ \hline 180 \end{array} \quad (8)$$

$$\begin{array}{r} 62 \\ +35 \\ \hline 97 \end{array} \quad (2)$$

$$\begin{array}{r} 62 \\ +35 \\ \hline 97 \end{array} \quad (2)$$

Scoring Template

$$\begin{array}{r}
 25 \\
 15 \overline{)375} \\
 \underline{30} \phantom{0} \\
 75 \\
 \underline{75} \\
 0
 \end{array}
 \quad (9)$$

Student Response

$$\begin{array}{r}
 25 \\
 15 \overline{)375} \\
 \underline{30} \phantom{0} \\
 75 \\
 \underline{75} \\
 0
 \end{array}
 \quad (9)$$

**Rule 5. Alignment for Incorrect Answers.** If the answer is incorrect, the digits must be correctly aligned to earn credit for each digit.

Scoring Template

$$\begin{array}{r}
 15 \\
 \times 12 \\
 \hline
 30 \\
 150 \\
 \hline
 180
 \end{array}
 \quad (8)$$

Student Response

$$\begin{array}{r}
 15 \\
 \times 12 \\
 \hline
 30 \\
 150 \\
 \hline
 180
 \end{array}
 \quad (6)$$

$$\begin{array}{r}
 62 \\
 + 35 \\
 \hline
 97
 \end{array}
 \quad (2)$$

$$\begin{array}{r}
 62 \\
 + 35 \\
 \hline
 67
 \end{array}
 \quad (1)$$

$$\begin{array}{r}
 25 \\
 15 \overline{)375} \\
 \underline{30} \phantom{0} \\
 75 \\
 \underline{75} \\
 0
 \end{array}
 \quad (9)$$

$$\begin{array}{r}
 26 \\
 15 \overline{)375} \\
 \underline{30} \phantom{0} \\
 75 \\
 \underline{70} \\
 5
 \end{array}
 \quad (2)$$

---

**Rule 6. Reversed Digits.** Reversed digits are counted as correct.

Scoring Template

$$\begin{array}{r} 22 \\ +32 \\ \hline 54 \end{array} \quad (2)$$

Student Response

$$\begin{array}{r} 22 \\ +32 \\ \hline 54 \end{array} \quad (2)$$

---

**Rule 7. Rotated Digits.** Rotated digits are counted as correct, with the exception of 6 and 9.

Scoring Template

$$\begin{array}{r} 61 \\ +25 \\ \hline 86 \end{array} \quad (2)$$

Student Response

$$\begin{array}{r} 61 \\ +25 \\ \hline 89 \end{array} \quad (1)$$

$$\begin{array}{r} 22 \\ +40 \\ \hline 62 \end{array} \quad (2)$$

$$\begin{array}{r} 22 \\ +40 \\ \hline 62 \end{array} \quad (2)$$

---

### Supplemental Rules for Multiplication and Division

---

**Rule 8. Longest Method.** For correct answers to division problems that are not basic facts, the student earns full credit for the "longest method" taught to solve the problem, even if the work is not shown.

Scoring Template

$$\begin{array}{r} 15 \\ 9 \overline{)135} \\ \underline{9} \phantom{0} \\ 45 \\ \underline{45} \\ 0 \end{array} \quad (8)$$

Student Response

$$\begin{array}{r} 15 \\ 9 \overline{)135} \end{array} \quad (8)$$

**Rule 9. Place Holders.** In multiplication, an "X," "0," or "(blank)," that counts as a place holder is scored as a correct digit.

Scoring Template

$$\begin{array}{r} 347 \\ \times 19 \\ \hline 3123 \\ 3470 \\ \hline 6593 \end{array} \quad (12)$$

Student Response

$$\begin{array}{r} 347 \\ \times 19 \\ \hline 3123 \\ 347 \times \\ \hline 6593 \end{array} \quad (12)$$

**Exception:**

When multiplying by a multiple of ten or hundred, only the digits in the answer are scored.

Scoring Template

$$\begin{array}{r} 122 \\ \times 300 \\ \hline 36600 \end{array} \quad (5)$$

Student Response

$$\begin{array}{r} 122 \\ \times 300 \\ \hline 000 \\ 0000 \\ 36600 \\ \hline 36600 \end{array} \quad (5)$$

**Rule 10. Remainders.** In division, remainders are scored as correct digits. Zero remainders are scored as correct digits, only once.

Scoring Template

$$\begin{array}{r} 11 \\ 10 \overline{)112} \\ \underline{10} \\ 12 \\ \underline{10} \\ 2 \end{array} \quad (9)$$

Student Response

$$\begin{array}{r} 11r2 \\ 10 \overline{)112} \\ \underline{10} \\ 12 \\ \underline{10} \\ 2 \end{array} \quad (9)$$

**Rule 11. Decimals** When calculation involves decimals, a decimal point must appear in the correct location in the answer, and is counted as a digit.

Scoring Template

$$\begin{array}{r} 1.65 \\ + .30 \\ \hline 1.95 \end{array} \quad (4)$$

Student Response

$$\begin{array}{r} 1.65 \\ + .30 \\ \hline 1.95 \end{array} \quad (3)$$

---

**Rule 12. Integers** When calculation involves integers, a positive or negative sign must appear in the correct location in the answer, and is counted as a digit. If the answer is positive, credit is still given if the (+) sign is not written.

<u>Scoring Template</u>	<u>Student Response</u>
$(-5) + (-2) = -7$ (2)	$(-5) + (-2) = -7$ (2)
$(+3) + (-2) = +1$ (2)	$(+3) + (-2) = 1$ (2)

---

**Rule 13. Horizontal format** When a problem is presented in horizontal format on a probe, only the answer is scored for correct digits.

<u>Scoring Template</u>	<u>Student Response</u>
$57 \div 18 = 39$ (2)	$\begin{array}{r} 39 \\ 18 \overline{) 57} \\ \underline{54} \phantom{0} \\ 39 \phantom{0} \end{array}$
$21 \times 40 = 840$ (3)	$\begin{array}{r} 21 \times 40 = 840 \\ \times 40 \\ \hline 840 \end{array}$

## Appendix F

### Inter-Rater Consent Form and Letter



March 21, 2001

Dear

I am a Resource Teacher at Heather Park Middle School, who is also a graduate student at UNBC working towards my Master of Education degree in Curriculum and Instruction. My thesis will investigate Gender and Relative-age Differences in Math Fluency Using Curriculum-Based Measurement. The thesis data will use the scores obtained during School District No. 57's CBM Math Norming Project in 1999-2000. Permission has been obtained from School District No. 57 to carry out this research.

Another aspect of the thesis will be a study of Inter-Rater Reliability (Marking reliability) of the CBM Math probes. Many different markers participated in scoring the probes. Validity of the CBM Math data will increase by determining the effect of a large number of markers scoring the math probes. As one of the people who attended the inservice training, your assistance is requested in the marking of 15 grade 7 math probes from the norming project. A package containing the 15 probes and a consent form will be sent to markers once their agreement is received. Identifying information has been removed from the Math probes including student names, scores and school. After the probes are marked please return them within two weeks, to BJ Foulds at Heather Park Middle School. Please include the signed consent form.

If you agree to assist in this research by participating in the Inter-Rater Study please inform BJ Foulds. The probes and signed consent form will be sent, once you notify me of your agreement to participate. Contact me either by phone, email or in person at the numbers below.

Phone- 962-1811 extension 614 (school) or 964-8267 (home)

Email- online 57 (BJ Foulds) or [fouldsb@unbc.ca](mailto:fouldsb@unbc.ca)

Please contact me if you have any questions regarding this study. If you require further information please contact the Thesis Supervisor, Dr. Peter MacMillan, The University of Northern British Columbia, telephone 960-5828 or by email at [peterm@unbc.ca](mailto:peterm@unbc.ca). If there are any complaints direct them to the Office of Research and Graduate Studies, UNBC.

Thank-you in advance for your co-operation.

Sincerely,

Bonnie-Jean (BJ) Foulds  
Resource Teacher  
Heather Park Middle School

**UNIVERSITY OF NORTHERN BRITISH COLUMBIA**  
**College of Arts, Social and Health Sciences**  
**Education Program**

**Master of Education Thesis**  
**Researcher: Bonnie-Jean (BJ) Foulds**

**Gender and Relative-age Differences in Math Fluency**  
**Using Curriculum-Based Measurement**

**Consent Form for CBM Math Probe Marker Participation**

---

A study in Inter-Rater reliability will be conducted as part of my Master's thesis. As one of the trained markers, you are requested to mark 15 grade 7 CBM Math probes from the School District No. 57 norming project of 1999-2000. School District No. 57 is aware of this request.

Before indicating your consent to participate in this study, it is required that you note your agreement to the following terms:

- I understand that all information received will be treated in an anonymous fashion and maintained in a secured location. Only the Thesis Supervisor will see the signed consent form and the probes from the Inter-Rater study. Upon receipt of the marked probes all identifying information will be removed. Identifying information including school district personnel names, and schools will not be used for the study. Upon completion of all research the scored probes and consent forms will be destroyed.
  - I understand that as a participant I am free to terminate participation at any time.
  - I understand that there is no remuneration for my participation in this study.
  - I understand that School District No. 57 has given permission for this research to proceed.
  - I understand that I may meet with the researcher to receive a verbal report of the findings when the thesis is completed.
  - I understand the if I require further information regarding the assignment, I may contact the Thesis Supervisor, Dr. Peter MacMillan, The University of Northern British Columbia, telephone 960-5828 or by email at [peterm@unbc.ca](mailto:peterm@unbc.ca). Direct any complaints to the Office of Research and Graduate Studies, UNBC.
- 

By signing this form I am providing written consent for participation of the Inter-Rater Reliability study investigating Gender and Relative-age Differences in the Math Fluency using the CBM Math norming data of School District No. 57 from 1999-2000.

Researcher: Bonnie-Jean (BJ) A. Foulds

Participant's Signature: \_\_\_\_\_ Date: \_\_\_\_\_

## Appendix G

### Table 1

#### Examples of Questions with Marker Discrepancies

Rater Probe #	Norming Probe #	Skill	Concerns/Comments	Scores	Range of Scores
1	4	2-digits x 2-digits	Answer incomplete, not all digits correct, #'s in row 1 are small and row 2 are large and on an angle	2, 1, 0	2
1	4	4-digit ÷ 2-digit	#'s difficult to read	1, 0, Unmarked	1
1	4	3-digit with decimals ÷ 1-digit	Answer correct, student used short-cut method, question partially completed	10, 3, 5	7
1	4	Conversion from decimal to percent	Whole # correct, decimal written as fraction	4, 2, 3, 0	4
1	4	Addition with decimals	Answer correct	6, 5	1
1	4	4-digit X 2-digit	1-digit incorrect, #'s small, maybe difficult to read, question complete	17, 16	1
2	5	3-digit X 2-digit with decimals	Some #'s may be difficult to decipher, decimal in wrong place, some digits incorrect	10, 9, 4, 3, 2	8
2	5	Integer subtraction	Negative sign correct but numeral incorrect	1, 0	1
2	5	Conversion from decimal to percent	Numerals correct, decimal in wrong place	3, 1, 0	3
2	5	2-digit X unknown # = 4-digit #	Answer correct	2, Unmarked	2
3	3	3 + 4-digits	Answer correct	4, Unmarked	4
3	3	4-digits X 2-digits	1 <sup>st</sup> row correct, did not insert placeholder for 10's in 2 <sup>nd</sup> row, completed question	9, 8, 6, 5, 4, 2	7
3	3	3-digits ÷ 1-digit with decimals	Work correct, only mistake is decimal incorrectly placed	9, 8, 1	8

3	3	Horizontal subtraction with decimals	Answer is not on the line provided, answer may be attempt to rewrite question since it repeats the first #, decimal in the correct place	3, 1, 0, Unmarked	3
3	3	Conversion from decimal to percent (3-digits in question)	No decimal in answer, decimal implied, only 2-digits for answer,	2, 1, 0, Unmarked	2
4	2	3-digit X 2-digit with decimals	Answer correct, decimal in correct location, blank used for a place holder	13, 12	1
4	2	4-digit ÷ 2-digit	Work incomplete, last 2 rows incorrect, 1 <sup>st</sup> digit of answer correct	11, 8, 7, 6, 3, 1	10
4	2	Percent of a number	Answer has 3-digits, should be 2, unnecessary decimal appears to be in the answer, numeral on right is correct	1, 0, Unmarked	1
4	2	Integer subtraction	Answer is positive, digit answer incorrect, positive sign in answer is not written (inferred)	1, 0, Unmarked	1
5	1	5-digits minus 4-digits	1-digit in answer incorrect	4, 3	1
5	1	Integer subtraction	Answer is positive, digit answer incorrect, positive sign in answer is provided	1, 0, Unmarked	1
5	1	3-digits ÷ 1-digit with decimals	No work shown, 4-digits in answer but should be 3, # after decimal is correct, 2 middle #'s are incorrect, # on left correct	3, 2, 1, 0	3
5	1	3-digits ÷ 1-digit with decimals	Work shown, alignment incorrect, some digits incorrect, no decimal in answer	8, 6, 5, 4, 3, 2	6
5	1	Conversion of improper fraction to whole #	Correct answer, only worth 1 mark	4, 1, Unmarked (Did rater copy the answer-4)	4
5	1	Integer multiplication	Answer correct	3, Unmarked	3

6	3	Unknown dividend $\div$ 2-digits = 2-digits	Answer a hundred instead of a thousand, digits correct except missing zero for one's	3, 1, 0	3
6	3	3-digits $\div$ 1-digit with decimals	Numerals in answer correct but decimal in wrong place, not all work shown Possible score=10	9, 6, 5, 4, 0	9
6	3	4-digits X 2-digits	Answer incomplete	8, 6, Unmarked	8
7	6	4-digit $\div$ 2-digit	Correct answer evident as in answer key, but student wrote remainder as decimal, did not stop with remainder answer, therefore extra work and numbers exist in answer from the answer key	10, 9, 8, 7, 5, 2	8
7	6	3-digit $\div$ 2-digit	Answer correct, work shown, alignment might be considered out for 1 or 2 numbers	9, 2	7
8	6	Integer subtraction	Negative sign correct, digits for answer incorrect	1, 0, Unmarked	1
8	6	2-digits X 2-digits	Answer correct, all work shown, 1 # written over another to correct a mistake	10, 4, Unmarked	10
8	6	4-digit $\div$ 1-digit	Answer correct with a decimal in the correct place but not required, no work shown	14, 13, 5, 3, Unmarked	14
9	4	Unknown quotient $\div$ 2-digit = 2-digit	Student answer a 2-digit #, ones digit correct, answer should be 4-digits	1, 0, Unmarked	1
9	4	Integer addition	Answer should be positive, digits incorrect, no sign (positive sign inferred)	1, 0, Unmarked	0
9	4	Integer division	Answer should be positive, positive sign inferred but numeral answer incorrect	1, 0, Unmarked	1
10	2	3-digits X 2-digits	All work shown, numbers very large, 2-digits incorrect in work, possible alignment concerns for 2 or 3 digits	12, 11, 9, Unmarked	12
10	2	Horizontal subtraction with decimals	One digit wrong in answer, decimal in correct place	4, 3, Unmarked	4

10	2	4-digit subtraction	3-digits correct, alignment concerns, a digit written over another	4, 3, Unmarked	4
10	2	5-digit X 1-digit	Hard to determine if the numeral "1", which is correct, is crossed out or not	3, 2, 0, Unmarked	3
10	2	Integer addition	Negative sign is above the numerals instead of to the left, otherwise answer is correct including the negative sign	3, 2, 1, 0, Unmarked	3
11	3	3 & 4-digit addition	One numeral incorrect in answer, numeral in thousand position may be out of alignment	4, 3, 2	2
11	3	Unknown $\div$ 2-digit = 2-digit	Answer should have 4-digits, student answer has 3 & is missing zero in ones	3, 1, 0	3
11	3	Integer addition	Student answer is zero with a possible small negative sign in the zero, answer should be negative	1, 0, Unmarked	1
11	3	4-digit X 2-digit	Two numbers rewritten over, shadow of previous number still visible, answer is correct	5, 4, 1	4
13	1	5-digits minus 4-digits	Student answer correct in tens & ones, digit in hundreds is correct number for thousands, correct # for hundreds missing	4, 3, 2	2
13	1	Integer addition	Answer should be positive, positive sign correct but digits incorrect	1, 0, Unmarked	0
13	1	2-digit X 2-digits	Answer incomplete, 2 of 3 digits correct	3, 2, Unmarked	3
13	1	Addition of 3 2-digit numbers	One of the 2-digits in the answer are correct	2, 1, Unmarked	2
13	1	Integer multiplication	Digits correct but negative sign should be positive	2, 1, Unmarked	2
15	4	4-digits X 2-digits	Answer incomplete, questions scribbled over but most #'s readable	8, 7, 6, 5, 4, Unmarked	8
15	4	Integer multiplication	Answer should be positive, no positive sign written, digits incorrect	1, 0, Unmarked	1

## Appendix H

### Table 1

Repeated-Measures ANOVA for the Average Performance Group



Grade	Source	F	df		p
1	Gender	0.56	1		.53
	RA	0.36	2		.73
	G*RA	1.17	2		.32
		F (from V)	df <sub>h</sub>	df <sub>e</sub>	p
2	Gender	0.59	2	115	.56
	RA	0.75	4	232	.56
	G*RA	0.41	4	232	.80
3	Gender	0.84	2	119	.44
	RA	2.58	4	240	.04
	G*RA	1.06	4	240	.38
4	Gender	1.44	2	113	.24
	RA	0.37	4	228	.83
	G*RA	0.93	4	228	.45
5	Gender	1.19	2	121	.31
	RA	1.30	4	244	.27
	G*RA	0.72	4	244	.58
6	Gender	0.15	2	117	.86
	RA	0.73	4	236	.58
	G*RA	1.44	4	236	.22
7	Gender	0.56	2	115	.58
	RA	0.36	4	232	.84
	G*RA	2.89	4	232	.02

$\alpha = .01$  for V and F